

RESEARCH ARTICLE

Molecular adaptation in Rubisco: Discriminating between convergent evolution and positive selection using mechanistic and classical codon models

Sahar Parto^{1*}, Nicolas Lartillot²

1 Department of Biochemistry and Molecular Medicine, Université de Montréal, Montreal, Quebec, Canada,

2 Laboratoire de Biométrie et Biologie Évolutive, Université Lyon 1, CNRS, UMR, Lyon, France

* sahar.parto@umontreal.ca



OPEN ACCESS

Citation: Parto S, Lartillot N (2018) Molecular adaptation in Rubisco: Discriminating between convergent evolution and positive selection using mechanistic and classical codon models. PLoS ONE 13(2): e0192697. <https://doi.org/10.1371/journal.pone.0192697>

Editor: Art F. Y. Poon, Western University, CANADA

Received: September 6, 2017

Accepted: January 29, 2018

Published: February 12, 2018

Copyright: © 2018 Parto, Lartillot. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data—including the sequence alignment and phylogenetic tree—are included within the paper and its Supporting Information files. All source code and data (diffsel branch) have been uploaded to GitHub and are available at the following link: <https://github.com/bayesiancook/coevol>.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Rubisco (Ribulose-1, 5-biphosphate carboxylase/oxygenase) is the most important enzyme on earth, catalyzing the first step of photosynthetic CO₂ fixation. So, without it, there would be no storing of the sun's energy in plants. Molecular adaptation of Rubisco to C4 photosynthetic pathway has attracted a lot of attention. C4 plants, which comprise less than 5% of land plants, have evolved more efficient photosynthesis compared to C3 plants. Interestingly, a large number of independent transitions from C3 to C4 phenotype have occurred. Each time, the Rubisco enzyme has been subject to similar changes in selective pressure, thus providing an excellent model for convergent evolution at the molecular level. Molecular adaptation is often identified with positive selection and is typically characterized by an elevated ratio of non-synonymous to synonymous substitution rate (dN/dS). However, convergent adaptation is expected to leave a different molecular signature, taking the form of repeated transitions toward identical or similar amino acids. Here, we used a previously introduced codon-based differential-selection model to detect and quantify consistent patterns of convergent adaptation in Rubisco in eudicots. We further contrasted our results with those obtained by classical codon models based on the estimation of dN/dS. We found that the two classes of models tend to select distinct, although overlapping, sets of positions. This discrepancy in the results illustrates the conceptual difference between these models while emphasizing the need to better discriminate between qualitatively different selective regimes, by using a broader class of codon models than those currently considered in molecular evolutionary studies.

Introduction

Rubisco (Ribulose-1, 5-biphosphate carboxylase/oxygenase) is an enzyme that catalyzes the major step in carbon fixation in all photosynthetic organisms. It is the most abundant protein on earth [1], as it encompasses up to 50% of soluble proteins [2] and 20–30% of total nitrogen

[3] in C3 leaves. During carbon fixation, Rubisco reacts with both CO₂ and O₂ as its substrate, with poor distinguishing ability. The carboxylase activity results in the incorporation of inorganic carbon into the metabolic C3 pathway, whereas the oxygenase activity boosts the photorespiration pathway. The latter prompts both energy consumption and CO₂ loss.

The evolution of C3 pathway goes back to 3 billion years ago when the atmosphere comprised high CO₂ and low O₂. In those conditions, photorespiration would have rarely happened. However, under present atmospheric conditions (lower CO₂ and higher O₂ concentration), photorespiration can represent a significant proportion of the enzymatic activity of Rubisco, such that the efficiency of photosynthesis can be dropped by 40% under unfavorable climates like hot and dry conditions [4]. As a result, some plants have developed an evolved improvement to C3 pathway called C4 photosynthesis as an adaptation to these changes in the environment [5,6].

About 85% of plants use C3 photosynthetic pathway, covering 78.4 million km² land area, whereas less than 5% are C4 plants, with global coverage of 18.8 million km² [7,8]. The rate of photosynthesis is different in these groups, being much more efficient in C4 plants than in C3 species. C4 photosynthesis mostly evolved as an adaptation to intense light, high temperature and aridity [9]. Hence, C4 plants dominate the grassland plants in harsh climates such as tropical, subtropical and warm regions [10].

The evolution of C4 plants from C3 ancestors consists of both anatomical and biochemical changes. These modifications allow C4 plants to concentrate more CO₂ around Rubisco, such that the oxygenase activity and the subsequent photorespiration are partially or completely repressed. The kinetics of Rubisco has been altered in C4 plants, leading to lower specificity and higher efficiency [11,12,13].

Interestingly, a relatively large number of independent transitions from C3 to C4 phenotype have occurred across monocots and eudicots. Each time, the Rubisco enzyme has been subject to similar selective pressure for tuning the tradeoff between substrate specificity and yield. As a consequence, C4 photosynthesis is an excellent model for convergent evolution at the molecular level in response to environmental changes [14]. In terms of applications, finding features of C4 plants and applying them to C3 plants such as rice, can be potentially used to increase crop yields [15,16]. Considering the above issues, understanding how selection acts on Rubisco in C4 plants compared to C3 ancestors can be very beneficial.

Based on these considerations, the evolution of Rubisco has attracted a lot of attention in recent years [17,18,19,20,21,22,23]. Kapralov et al [19] tried to identify positive selection in Rubisco using molecular phylogenetic analyses. Employing codon models that allow for varying selection among sites (implemented in CodeML [24]), they detected sites under positive selection in some photosynthetic organisms, especially in the main lineages of land plants. More recently, Kapralov et al [20] used a similar method to investigate the evolution of Rubisco in C4 plants in a large group of C4 eudicots and found sites under positive selection. They observed that some of those positively selected sites appear to display consistent patterns of amino acid substitutions associated with the C3 to C4 transition.

These empirical analyses raise an interesting question, concerning the use of codon models to characterize selective regimes in protein-coding sequences. Typically, elevated dN/dS results from ongoing adaptive processes, by which a protein-coding gene is constantly challenged by ever-changing selective forces. However, in its general form, this process of ongoing adaptation is not associated to repeated transitions toward the same amino acid at a given position, independently across multiple lineages, and could instead constantly elicit new amino acids at positively selected sites. In contrast, the multiple transitions between C3 and C4 photosynthetic regimes represent a case of convergent evolution. At the molecular level, this is expected to result in recurrent directional selection, thus, potentially favoring the same amino acid(s) at

the same site(s) upon each C3 to C4 transition. In addition, the overall dN/dS induced by this process of recurrent directional selection is fundamentally determined by the rate of C3/C4 transitions across the phylogeny, which may not be sufficiently high to induce a dN/dS greater than 1 at those positions that are susceptible to respond to this convergent evolutionary process. Thus, positive selection, like what is formalized by classical codon models (i.e. by an elevated dN/dS), may not be the most appropriate selective regime to test in the present case. A similar distinction between episodic diversifying and directional selection has been previously proposed by Murrel et al [25]. They demonstrated that modeling the episodic and directional selection explicitly enhance the accuracy to identify drug-resistant sites in HIV-1. More recently, Thiltgen et al [26] compared two directional selection models (MEDS [25] and swMutSel [27]) with each other and with the standard method of detecting diversifying selection (PAML), on the same dataset as [25]. None of the three models could outperform the others in their study.

Convergent amino acid substitutions which potentially linked to adaptation to the C4 phenotype have been more directly investigated by Studer et al [23]. These authors used the TDG09 model, allowing for site- and condition-specific amino acid preferences [28], to identify sites under condition-dependent selection. Recently, we have developed an approach similar to the TDG09 model, called Differential Selection (DS) model [29] using a Bayesian mechanistic derivation of the codon substitution process, under the so-called mutation-selection formalism. Here, we re-assessed the question of positive versus convergent selective patterns in the Rubisco gene in eudicots, using two types of codon models: first, we applied our DS model to identify amino acids which are differentially selected at specific positions along the Rubisco sequence, as a function of the photosynthesis pathway. Second, we implemented Bayesian versions of the classical dN/dS-based codon models, allowing for both site- and branch-specific modulations of the dN/dS ratio [30], and applied them to the Rubisco dataset. We found that the two classes of models tend to select distinct, although overlapping, sets of positions. Altogether, our analysis emphasizes the existence of qualitatively different adaptive regimes undergone by protein-coding genes, and the need to better discriminate between these distinct regimes by using a broader class of codon models than those currently considered in molecular evolutionary studies.

Materials and methods

Sequence data, phylogenetic tree, and partitioning scheme

We obtained the *Amaranthaceae* rbcL multiple sequence alignment and the original phylogenetic tree (Fig 1) from Kapralov et al [20]. The dataset consists of 179 rbcL sequences of length 1341 base pairs, corresponding to amino acids 22–468 (the first 21 coding positions are missing). Out of 179 sequences, 84 and 95 sequences belong to C4 and C3 species, respectively. The list of species and their photosynthetic type (C3 or C4) is shown as supplementary information; S1 Table.

The phylogenetic tree was partitioned according to two alternative schemes, with $K = 3$ or $K = 2$ distinct conditions, based on the type of the photosynthetic pathway. In the three-condition scheme, the largest monophyletic clades exclusively composed of C3 or C4 species were first identified and defined as conditions 1 and 2. The branches at the base of each C3 and C4 clades were also included in conditions 1 and 2, respectively. All other branches outside from these clades (reconstructed ancestral branches) were considered as belonging to condition 0. The model that employs this approach is called DS3 and its phylogenetic tree is illustrated in Fig 1. The two-condition setup (model DS2) differs from the three-condition scheme (model DS3) by allocating all branches outside of the C4 monophyletic clades (together with their

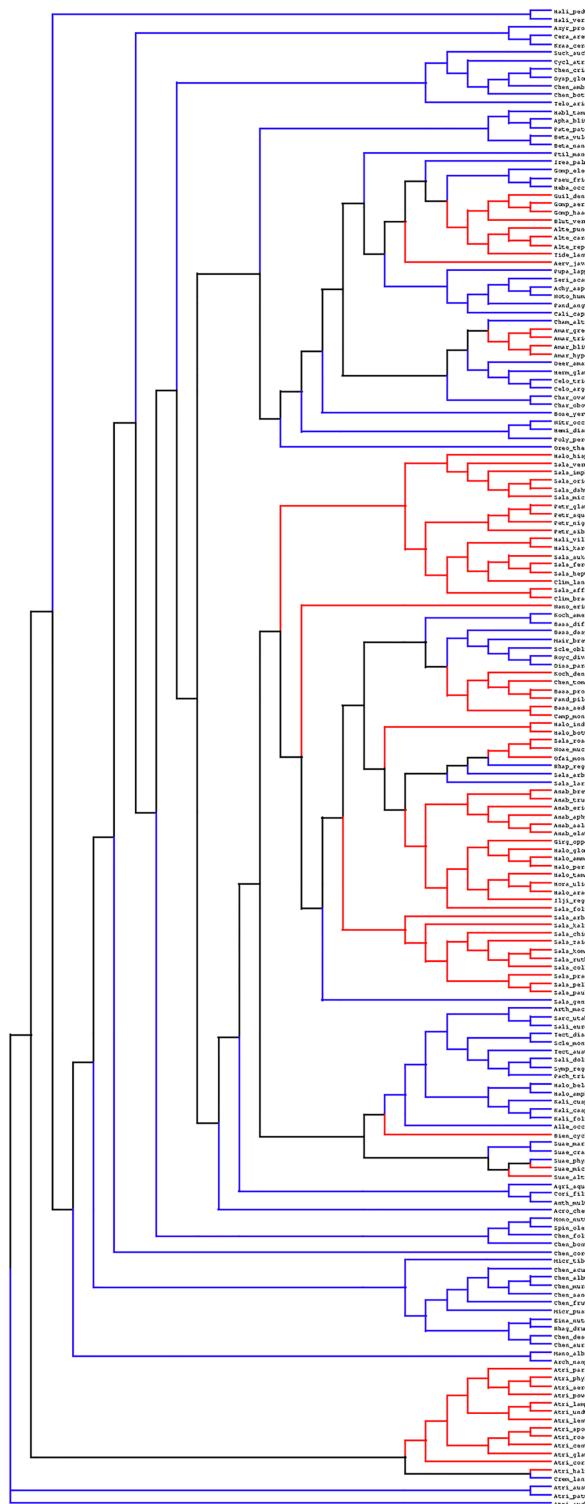


Fig 1. Phylogenetic tree of 179 rbcL sequences from Amaranthaceae family. The tree partitioned (according to model DS3) in C3 (blue), C4 (red) and interior branches (black). Number on each branch is the bootstrap support (provided by [20]). The tree is visualized using Dendroscope program [31].

<https://doi.org/10.1371/journal.pone.0192697.g001>

basal branches) by default to the C3 condition. Model DS2 amounts to assuming a maximum-parsimony reconstruction of the evolution of the photosynthetic regime, under the assumption that evolutionary transitions are exclusively from C3 to C4, with no reversion back to C3 [32]. However, model DS2 statistically implies a comparison between two conditions that are unevenly represented along the phylogeny, both in terms of total number of branches (152 for C4 versus 203 for C3) and concerning the evolutionary depth (the DS3 condition is mostly represented by recent branches, while the DS2 condition encompasses both ancient and recent lineages). In this respect, the advantage of model DS3 is to balance the empirical signal between the two conditions of interest (C3 and C4, represented by 158 and 153 branches under the DS3 model), and to focus exclusively on recent branches of the phylogeny for both conditions.

Differential-selection model

The principles of the differential-selection model were introduced previously [29] and we only recall the general structure here.

We used mutation-selection formalism, as in Halpern and Bruno [33] or Rodrigue et al [34]. According to this formalism, the substitution rates between codons were derived from first principles of population genetics, in terms of mutation rates and selective effects. The latter was explicitly modeled and assumed to operate exclusively at the level of the amino acid sequence.

More specifically, consider a sequence of N coding positions ($3N$ nucleotide positions). The number of conditions across the phylogenetic tree is denoted as K ($K = 2$ or $K = 3$, depending on the partition scheme). The mutation process is assumed to be time-reversible and homogeneous among sites and across lineages. It is thus entirely characterized by a general time-reversible 4×4 matrix Q . In contrast, the selective forces acting at the amino acid level are both condition- and position-specific. Accordingly, for each position, $i \in [1, N]$ and each condition $k \in [1, K]$, we introduced an array of 20 non-negative fitness factors, $F^{ik} = (F_a^{ik})_{a \in [1, 20]}$, one for each amino acid. In the following, these 20-dimensional vectors will be referred to as amino acid *fitness profiles*. In the present version of the model, they are assumed to be random effects across sites and conditions, drawn *iid.* from a uniform Dirichlet distribution.

Once these mutation rates and fitness factors are specified, the substitution process can be defined as follows. Consider the substitution rate between codon c_1 to c_2 (encoding amino acids a_1 and a_2) at site i and condition k , where codons c_1 and c_2 are assumed to vary only at one nucleotide position, with respective nucleotide states n_1 and n_2 at that position. First, we defined a Darwinian scaled selection coefficient, associated with a mutation from wild-type codon c_1 to mutant codon c_2 . Since selection is assumed to act only at the level of the amino acid sequence, this scaled selection coefficient is given by

$$S_{a_1 a_2}^{ik} = \ln \left(\frac{F_{a_2}^{ik}}{F_{a_1}^{ik}} \right)$$

Then, the rate of substitution between codon c_1 and c_2 is given by the product of the mutation rate and the relative fixation probability P (i.e. relative to neutral). This fixation probability is itself dependent on the scaled selection coefficient just defined. Using the classical diffusion approximation, this relative fixation probability is expressed as

$$P_{fix} = \frac{S_{a_1 a_2}^{ik}}{1 - e^{-S_{a_1 a_2}^{ik}}}$$

Thus, finally, the rate of substitution between codons c_1 and c_2 at position i and under condition k is given by

$$\text{Rate}_{c_1 c_2} = \begin{cases} Q_{n_1 n_2} & S_{a_1 a_2}^{ik} \\ Q_{n_1 n_2} \times \frac{S_{a_1 a_2}^{ik}}{1 - e^{-S_{a_1 a_2}^{ik}}} & 0 \end{cases}$$

Omega-based codon models

As an alternative to mutation-selection models, one of the most well-known and widely used methods for characterizing the selective regimes, involved in the evolution of protein-coding genes, is to estimate the ratio of non-synonymous (dN) to synonymous (dS) substitution rate (dN/dS), denoted as ω . These omega-based models were first proposed by Goldman and Yang [35] and Muse and Gaut [36], and subsequently complexified to account for site- and branch-specific modulations of the dN/dS ratio [30,37,38,39].

Here, we used the Muse and Gaut formalism and proposed a Bayesian model allowing for site- and condition-specific modulations of $\omega = \text{dN}/\text{dS}$. According to this model, the instantaneous substitution rate from codon c_1 to c_2 at site i and condition k is specified as follows

$$\text{Rate}_{c_1 c_2} = \begin{cases} Q_{n_1 n_2} & \omega^{ik} \\ Q_{n_1 n_2} \times \omega^{ik} & 0 \end{cases}$$

synonymous substitution
non-synonymous substitution
multiple nucleotide replacement

Here, ω^{ik} is thus the dN/dS ratio for site i and under condition k . For each condition k , the ω^{ik} 's, for $i \in [1, N]$ are modeled as random effects across sites, drawn *iid* from a gamma distribution of shape and scale parameters α^k and β^k .

We considered two alternative versions of this omega-based model: in model OM1, we assumed only one condition, thus defining a single (global) value of ω^i across the whole phylogenetic tree for site i ; in model OM3, on the other hand, the tree is partitioned into three conditions according to the photosynthesis pathways, exactly as for model DS3 above, and a distinct value ω^{ik} is allowed for site i and under condition $k \in [1, 3]$.

Priors. In all analyses presented below, the topology (τ) of the tree is fixed. For all models, the prior on branch lengths is a product of independent Exponentials of mean λ ; the hyper-parameter λ is from an Exponential distribution of mean 0.1; the prior on relative exchangeabilities of the mutation process is a product of Exponentials of mean 1; the prior on the mutational equilibrium frequency vector is a uniform Dirichlet distribution. As mentioned above, under the DS2 and DS3 models, the site- and condition-specific fitness profiles, $F_{a_1 a_2}^{ik}$, are random effects integrated over a Dirichlet distribution. Concerning the OM1 and OM3 models, the site- and condition-specific dN/dS values (ω^{ik}) are random effects integrated over a gamma distribution of shape and scale parameters α^k and β^k , which are themselves drawn from an exponential prior of mean 1 for each $k \in [1, K]$.

MCMC sampling

To sample the parameters from their joint posterior distribution, we used the general MCMC approach previously described in [29,40,41]. This approach consists of an alternation between stochastic mapping of the detailed substitution history at each coding site, followed by a long series of Metropolis-Hastings updates of all parameters and all random effects across sites and across conditions, conditional on this stochastic mapping.

Two independent MCMC were run for each analysis. In all cases, burn-in was first estimated visually, and then convergence and mixing were quantified using the tracecomp program (from the Phylobayes suite [42]) to compare the samples obtained under independent runs. Tracecomp gives an estimate of the discrepancy between the two runs, as well as the effective sample size, for several key parameters and statistics of interest. In the present case, the minimum effective size was always greater than 3000 and the discrepancy less than 0.2 for most statistics. Finally, the reproducibility of the estimation of the posterior mean differential selection factors across all amino acids and all sites was verified by plotting the estimates for all amino acids and all sites across the two independent runs (S1 Fig). After 400 points of burn-in from a total of almost 6000 points have been removed, posterior estimates were obtained by averaging over the remaining of the MCMC run.

Post-analysis

Under the DS models, for a given configuration of the model (typically drawn from the posterior distribution by MCMC), Differential Selection between two conditions C3 and C4 is simply calculated as the log-ratio between the amino acid fitness profiles ascribed to conditions 1 and 2

$$D_{a_1 a_2}^i = \ln \left(\frac{F_{a_1 a_2}^{i2}}{F_{a_1 a_2}^{i1}} \right)$$

These arrays of 20 differential selection effects (for the 20 amino acids) at each position are then averaged over the posterior distribution by MCMC. A position is deemed to show strong statistical support for a differential effect in favor of amino acid a_2 (in condition C4) over amino acid a_1 (in condition C3) if the posterior probability that $D_{a_1 a_2}^i > 0$ is greater than 0.90. Conversely, strong support for a negative differential effect (i.e. a differential effect against a_2 in favor of a_1) is called whenever the posterior probability that $D_{a_1 a_2}^i < 0$ is greater than 0.90.

Under the OM models, the posterior mean value of site- and condition-specific dN/dS is reported. Position i is regarded to have a strong support for positive selection under condition k if the posterior probability that $\omega^{ik} > 1$ is greater than 0.90.

Results and discussion

Amaranthaceae is one of the plant families with the largest number of C4 species. This makes it a suitable case for Differential Selection (DS) analysis. Based on a multiple sequence alignment of rbcL genes and an annotated phylogenetic tree of *Amaranthaceae*, our DS model captures site-specific amino acid preferences as vectors of 20 fitness factors (for the 20 amino acids) under each condition. Then, contrasting for each position, the fitness factors estimated in the two conditions of interest (here, in the C3 and C4 regimes), allows us to identify positions for which the fitness of a specific amino acid has undergone a significant change, either upward or downward, associated with the transition between the C3 and the C4 photosynthetic regime (see Methods).

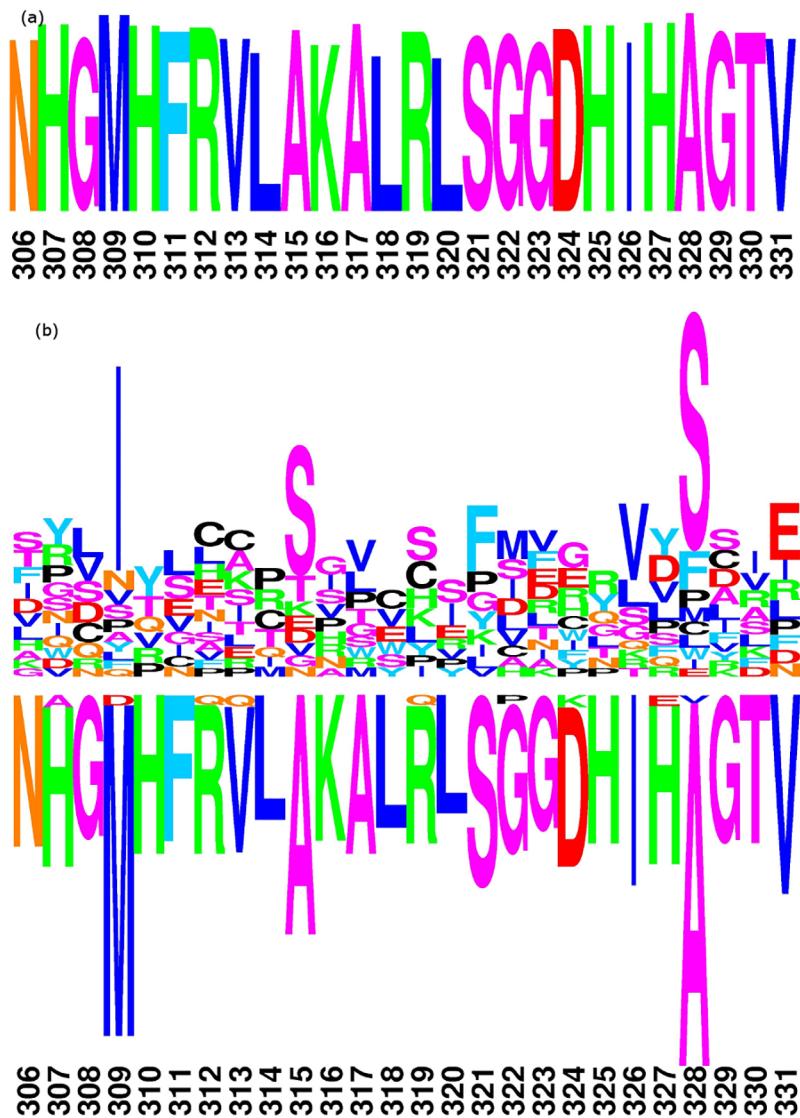


Fig 2. Global and differential selection profiles for position 306–331, under model DS2. (a) Amino acid fitness for C3 plants. (b) Differential amino acid fitness for C4 plants. Amino acids above (under) the line show an increase (decrease) in fitness compared to (a).

<https://doi.org/10.1371/journal.pone.0192697.g002>

DS2 versus DS3: Finding an optimal contrast between C3 and C4

As described in the methods, the tree was divided into either two or three conditions, resulting in two versions of the Differential Selection codon model, referred to as DS2 and DS3. In DS2, the interior branches (black branches in Fig 1) which connect C3 and C4 clusters are defined as C3. This corresponds to a plausible reconstruction of ancestral photosynthetic regimes across the group, as no reversal from C4 to C3 is known [32].

The selection profiles at position 306–331 estimated by DS2 are shown in Fig 2. In this Figure, we use a graphical logo representation [43] to display both absolute (global) and differential fitness distributions. Absolute logos for the reference condition represent the fitness of amino acids under a specific condition, with the height of the letter being proportional to the fitness of the corresponding amino acid. Differential logos, on the other hand, represent the

difference in log-fitness between two conditions: letters above (resp. below) the baseline corresponds to amino acids whose fitness is increased (resp. decreased) in each condition, compared to its parent condition.

The global selection profile (Fig 2-a) captures the absolute amino acid fitness for C3 plants. This profile primarily reflects the strong conservation of the protein sequence, with one single amino acid overwhelmingly favored at most positions. The differential profile between C4 and C3 (Fig 2-b) shows interesting patterns of opposite selective effects concerning pairs of amino acids, specifically at positions 309, 315 and 328. However, the differential profile between the C4 and C3 is also characterized by an inferred background of apparently non-specific differential selective effects concerning all major amino acids represented in the absolute fitness profile under C3: essentially, the absolute profile under C3 displays the consensus sequence of the alignment, while the differential profile between C4 and C3 reproduces this consensus sequence, although now below the line. This is likely to be a statistical artifact, which might have two alternative explanations. The first one is **the possible existence of non-fixed polymorphic states in the multiple sequence alignment**. These mutations, whose fate is to be ultimately removed by purifying selection, are expected to be mapped specifically along the terminal branches of the phylogeny, and may thus contribute to an apparent decrease in the inferred fitness of ancestral amino acids in the condition that is most enriched in terminal branches (here, C4). Another possible explanation is that **the number of branches allocated to the C4 condition is smaller than that allocated to the C3 condition, potentially leading to a difference in statistical power between the two conditions**. As a result, and in the presence of shrinkage mediated by the prior, the fitness of conserved amino acids is inferred to be higher in that condition that is endowed with the largest number of branches (here, C3).

One way to avoid this artifact is to balance the signal between condition C3 and C4, by allocating the interior branches of the tree to another baseline condition and by restricting the inference of C3-specific selection to the monophyletic groups of C3 species. In the present case, there are a comparable number of C3 and C4 branches (about 150 for each). This new setting (model DS3) is therefore expected to result in a much more balanced assessment of the differential-selection effects between recent C3 and C4 lineages.

Indeed, and unlike the differential profile between C3 and C4 provided by the DS2 model (Fig 2), the differential profile given by DS3 between recent C3 and C4 lineages (Fig 3) appears to have more reasonable properties: sparse, selecting a small number of positions for which specific amino acids appear to be differentially selected between the two photosynthetic regimes, and balanced between positive and negative effects (above and below the line, respectively). **For instance, at position 309, the fitness of Methionine is substantially decreased in C4 plants, compared to C3 species ($pp = 0.93$). Correlatively, the fitness of Isoleucine is increased at that position ($pp = 0.87$)**. Similarly, residue 328 is identified by the DS3 model as a position of the rbcL gene under the highest differential selection effect between C3 and C4 *Amaranthaceae* species. At site 328, Alanine is globally preferred in *Amaranthaceae* eudicots, yet in C4 group, its fitness is significantly decreased ($pp = 0.99$) in favor of Serine, whose fitness is increased compared to what prevails in C3 lineages ($pp = 0.96$). Based on these observations, in the following, we conduct all Differential Selection analyses under the DS3 model. The complete C4/C3 differential logo, for the whole sequence alignment, is displayed in the supplementary material (S2 Fig).

Differential selection (DS) versus omega-based (OM) codon models

 In addition to DS3, which belongs to the family of mutation-selection codon models (Halpern and Bruno [33] style), the *Amaranthaceae* dataset was also analyzed under two omega-based

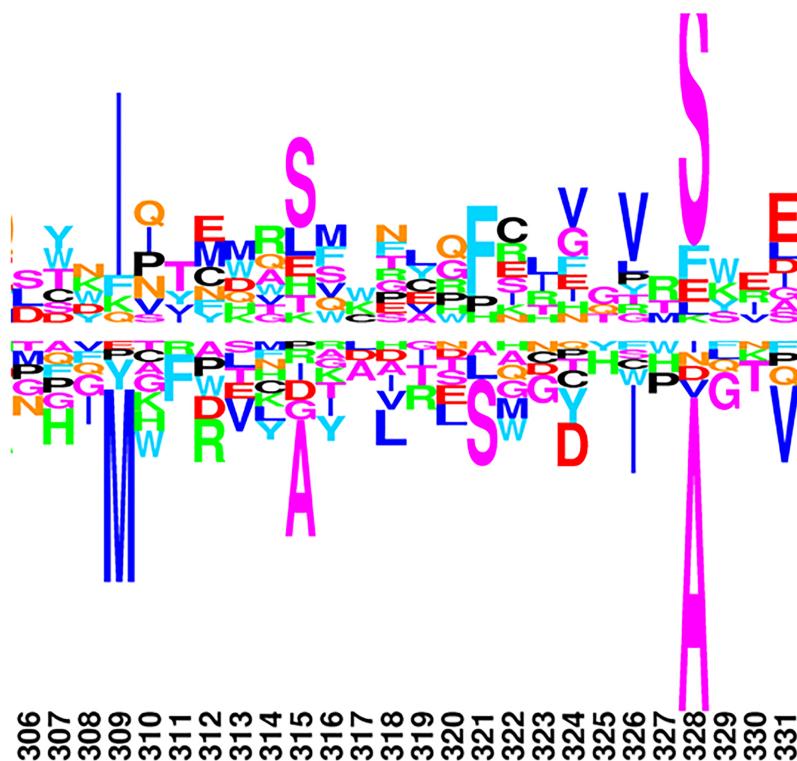


Fig 3. C4/C3 differential selection profile for position 309–328, under the DS3 model.

<https://doi.org/10.1371/journal.pone.0192697.g003>

models (Muse and Gaut [36] style), which we refer as OM models. The first of these models (OM1) is a site-specific model: each site has its own value for $\omega = dN/dS$, all of which are modeled as site-specific gamma-distributed random effects. By selecting sites with a high posterior probability of having a value of dN/dS greater than 1, model OM1 allows for the detection of sites under positive selection globally across *Amaranthaceae* dataset. The second model (OM3) allows for independent values of dN/dS , simultaneously across sites and conditions. Conditions are defined as in DS3 model (internal branches, as well as terminal C3 and C4 clades).

Thus, OM3, unlike OM1, allows for the detection of sites under positive selection specifically in C3 or in C4 species. To facilitate the comparison, all three models, DS3, OM1 and OM3, were implemented in a Bayesian framework, using similar strategies for designing the models (both dN/dS and differential selective effects modelled as either global or condition-dependent *iid* random effects across sites) and for detecting significant effects (based on the posterior probability for a site to have a value of $\omega > 1$ or a differential selective effect greater and smaller than 0, globally or in a given condition). The results of these analyses are summarized in Table 1. In this table, all sites for which a strong support ($pp > 0.90$) was found under at least one of the three models are reported.

Under model OM1, 6 positions (32, 43, 145, 225, 262 and 279) were found to have a $dN/dS > 1$ with a posterior probability greater than 0.90, and 2 positions (439 and 443) with $pp < 0.90$. These 8 positions are exactly those reported by Kapralov et al [20], found using the BEB approach implemented in CodeML. Note that the approach used here and the one implemented in CodeML are rather different in their statistical strategy for detecting sites under positive selection. The approach of CodeML relies on a mixture model, whereas the present approach explicitly assumes independent values of dN/dS across sites. The results obtained

Table 1. Findings of OM1, OM3 and DS3 model. Only positions with posterior probability >0.9 in any of the above models are reported here. Positions specified with an asterisk are those found previously by Kapralov et al [20] (one for C3 and two for C4). ω^1 , ω^2 , and ω^3 represent ω values for condition 1, 2 (C3) and 3 (C4). The sign next to each amino acid shows the direction of selection.

Position	OM1 model		OM3 model						DS3 model	
	ω	pp($\omega > 1$)	ω^3	pp($\omega^3 > 1$)	ω^2	pp($\omega^2 > 1$)	ω^1	pp($\omega^1 > 1$)	Amino acid	pp
32*	3.2	0.99	1.32	0.61	3.68	0.99	3.76	0.97	+Q, -L, +K	0.93, 0.91, 0.81
43*	2.13	0.99	1.03	0.48	2.32	0.98	2.93	0.93	-	-
86	0.71	0.15	0.02	0	1.23	0.65	0.48	0.1	+H, -N	0.92, 0.89
143	0.55	0.05	0.01	0	0.9	0.34	0.43	0.14	+S, -A	0.94, 0.77
145*	2.65	0.99	3.4	0.99	2.14	0.96	0.06	0.01	-	-
225*	2.53	0.99	1.27	0.58	2.7	0.99	3.55	0.98	-L, +I	0.96, 0.88
262*	2.25	0.99	0.33	0.05	3.61	0.99	1.69	0.64	+V, -A	0.99, 0.75
279*	2.19	0.99	2.21	0.98	2.13	0.98	1.28	0.51	-	-
281**	1.11	0.62	2.44	0.99	0.33	0.02	0.28	0	-A, +S	0.96, 0.70
309**	0.4	0.006	1.08	0.47	0.01	0	0.02	0	-M, +I	0.94, 0.87
328	1.35	0.8	1.77	0.89	0.87	0.3	0.56	0.2	-A, +S	0.99, 0.98
354	0.77	0.21	0.02	0	1.43	0.72	0.03	0	+T, -I	0.92, 0.89
439*	1.15	0.63	0.15	0.01	2.05	0.98	0.26	0.08	-T, +R	0.89, 0.84
443*	1.5	0.88	0.01	0	2.28	0.99	1.79	0.72	+T, -A	0.97, 0.77
461	1.36	0.79	0.09	0.01	1.65	0.85	2.93	0.93	+V, -I	0.98, 0.86

<https://doi.org/10.1371/journal.pone.0192697.t001>

here, therefore, suggest that at least in the present context, the details of the overall statistical strategy do not have a strong influence on the outcome of the model.

Kapralov et al [20] also reported an additional two sites, 281 and 309, detected by the branch-site model and thus inferred to be under positive selection specifically in C4 condition. Here, using the OM3 model, which allows for condition- and site-specific values for dN/dS, we found statistical support for positive selection in C4 only for position 281. For position 309, the posterior mean value of omega in condition C4 is indeed greater than 1 (1.08), although only with a weak posterior probability support (pp = 0.47). Conversely, it is worth noting that several sites (such as 32, 43, 225, and 262) are inferred by model OM3 to be under positive selection only under C3, but not under C4.

Finally, the DS3 model uncovers a series of 11 sites under Differential Selection between C3 and C4 with pp > 0.90. These 11 sites include 4 of the sites discovered by model OM1, thus inferred to be under global positive selection (32, 225, 262 and 443), as well as sites 281 and 309 (inferred to be under positive selection specifically in C4, either by model OM3 or by branch-site models of CodeML). Conversely, and importantly, half of the discoveries made by the DS3 model (6 sites out of 11, including site 309) do not show any signal of positive selection under either OM1 or OM3.

Differential selection patterns in Amaranthaceae family

Here we studied the molecular adaptations associated with the C3 to C4 transition in *Amaranthaceae* eudicots. Using a mechanistic codon model for detecting differential selection patterns associated with these adaptations, we found 11 positions to be under Differential Selection pressure between C3 and C4 eudicots. Some of the amino acid substitutions undergone by these positions might have a conformational or catalytic role in Rubisco enzyme in C4 plants, leading to its higher efficiency [23,44]. Alternatively, they might be a compensatory mutation selected to maintain its optimized function.

For instance, residue 328, with the highest differential effect, locates in the active loop 6 of the enzyme. Replacement of hydrophobic A with polar S destabilizes the active site, which leads to more flexibility of its opening and closing [45] and might explain the higher efficiency of C4 plants. Site 281 lies in the core of C-terminal domain, and it may have a long-range effect on active loop 6 [23]. Position 309 is in the interface of C-terminal domains of two subunits within a dimer [23], which might have an effect on flexibility. Although residues 86, 354 and 461 are found to be under strong Differential Selection pressure between C3 and C4 *Amaranthaceae*, their exact role has not been specified. Position 461 locates near a large subunit residue (residue 466) which might account for the interaction with Rubisco activase.

Comparing differential selection and omega-based codon models

Previously, some positions have been found by other phylogenetic methods to be under specific selective regimes, potentially associated with the C3 to C4 transition. In particular, Kapralov et al [20] used the concept of dN/dS as selection strength along the coding sequence. Using classic dN/dS codon models, they uncovered a set of 10 positions putatively under positive selection, either globally over the tree (8 positions) or specifically in the C4 groups (2 positions). In order to further explore this point, we implemented new dN/dS codon models, allowing for site- and condition-specific dN/dS, in our Bayesian framework. Selecting sites based on the posterior probability support for $dN/dS > 1$, we essentially recovered the same set of positions as that reported by Kapralov et al (except for one position). On the other hand, if we compare the set of findings under dN/dS models and the Differential Selection model, we observe a partial overlap. Specifically, only half of the positions inferred to be under Differential Selection between C3 and C4 were also found by dN/dS models. Conversely, 4 of the 10 findings under both classes of dN/dS models showed differential selection effects.

 The partial overlap between the findings of omega-based and Differential Selection models illustrates the conceptual difference between these models and the fact that they are meant to capture fundamentally different selective patterns. Classic omega-based codon models are meant to detect an overall *acceleration* of the rate of non-synonymous substitution. Such accelerations are typically caused by *ongoing adaptation*, due to diversifying selection, ecological red-queens, or fluctuating selection caused by environmental changes. In contrast, Differential Selection models are intended to capture convergent patterns of *directional selection* associated with a specific change in the environment, having occurred several times independently across the phylogeny.

These two classes of selective patterns are not completely mutually exclusive. In principle, recurrent substitution events due to directional selection caused by repeated transitions from C3 to C4 photosynthesis across the *Amaranthaceae* family could result in an overall increase in  dN/dS observed at the corresponding sites. However, if the rate of C3 to C4 transitions is not sufficiently high, the resulting increase in dN/dS may not be enough to lead to a situation where $dN/dS > 1$. As a result, some of the important condition-specific adaptations might be missed by dN/dS codon models. For instance, as illustrated here, positions 86, 143 or 354, which show a strong differential-selection effect, yet have a dN/dS not exceeding 1.

In addition, this phenomenon of recurrent directional selection linked to repeated C3 to C4 transitions cannot explain that most of the sites inferred to be under positive selection have a $dN/dS > 1$ globally over the tree, and often (e.g. positions 32, 43 and 279) even specifically within the C3 terminal clades, in which no such substitution event induced by C3 to C4 transition is supposed to have occurred. Concerning positions 43 and 279, for instance, no differential selection effect is detected by the DS3 model, while the dN/dS is inferred to be of the order of 2, including within terminal C3 clades. Thus the most likely explanation for the pattern of

Darwinian evolution at those sites is simply the presence of ongoing adaptation that would not be directly related to the repeated transitions between C3 and C4 photosynthetic regimes.

Conversely, we observed some positions (in particular 262 and 461) that show a differential selection effect between C3 and C4, combined with a pattern of positive selection over the tree, except in the C4 condition, in which the dN/dS is specifically and markedly decreased (posterior mean dN/dS = 0.33 and 0.09, respectively). A possible explanation for this pattern is that, in C3 species, those positions are available for ongoing adaptation to a constantly fluctuating environment, but the transition to C4 photosynthesis essentially locks those positions into more specific adaptive amino acid states, thereby stopping the flux of adaptive substitutions at those sites. Of note, this concurrence of positive selection and differential selection effects (e.g. positions 32, 225) is not so easily explained in the context of the mutation-selection modeling framework used here. Mutation-selection models predict that the dN/dS should always be less than 1 at mutation-selection balance [46]. In the present case, this means that the DS model does not predict dN/dS greater than 1, except possibly during the transient phases following a change between the C3 and the C4 regimes—thus, at any rate, not within the C3 clades.

Conclusions

Rubisco has long been known to be under positive selection [19]. In addition, it has been shown that Rubisco has been evolved in different structural forms and functions [47]. It exemplifies a convergent evolution of enzyme properties in its phylogenetic pathway. One example of this convergent evolution happens between C3 and C4 plants through crossing the fitness landscape [20,23,48]. Therefore, the complex molecular evolutionary patterns displayed by the Rubisco gene in eudicots represent an interesting case-study for assessing and comparing current codon modeling strategies [20]. In this respect, our comparative analysis, by making an inventory of the amino acid-positions in rbcL sequences that are positively or differentially selected in C3 and C4 *Amaranthaceae* family, emphasizes the fundamental difference, in scope and meaning, between the two main classes of models currently considered in the literature: on one side, classic codon models based on the measure of the overall dN/dS, whose focus is primarily on positive selection; and on the other side, Differential Selection models, whose aim is instead, to detect convergent patterns of directional selection associated with repeated transitions between known evolutionary regimes. Our analysis also emphasizes that none of the models considered here, either omega-based models or mutation-selection approaches, offers a completely satisfactory explanation of the complex patterns of molecular evolution observed in *Amaranthaceae*, and probably also present in other species groups—thus suggesting that further developments are still needed on the front of phylogenetic codon models.

Supporting information

S1 Fig. Estimates of posterior mean differential selection effects across all amino acids and all sites for two independent chains, for C3 plants (a) and C4 plants (b).

(DOCX)

S2 Fig. C4 differential sequence logo for rbcL sequence in *Amaranthaceae* family.

(DOCX)

S1 Table. List of 179 species from *Amaranthaceae* family.

(DOCX)

S2 Table. Alignment of 179 sequences.

(TXT)

S3 Table. Phylogenetic tree.
(NWK)**Acknowledgments**

We are greatly thankful to Maxim Kapralov for sharing his dataset with us and for allowing us to make the data available online. We also thank the anonymous reviewers for their comments on the manuscript.

Author Contributions

Conceptualization: Sahar Parto, Nicolas Lartillot.

Methodology: Sahar Parto, Nicolas Lartillot.

Supervision: Nicolas Lartillot.

Writing – original draft: Sahar Parto.

Writing – review & editing: Nicolas Lartillot.

References

1. Ellis RJ The most abundant protein in the world. *Trends in Biochemical Sciences* 4: 241–244.
2. Feller U, Anders I, Mae T (2008) Rubiscolytics: fate of Rubisco after its enzymatic function in a cell is terminated. *J Exp Bot* 59: 1615–1624. <https://doi.org/10.1093/jxb/erm242> PMID: 17975207
3. Makino A (2003) Rubisco and nitrogen relationships in rice: Leaf photosynthesis and plant growth. *Soil Science and Plant Nutrition* 49: 319–327.
4. Ehleringer JR, Sage RF, Flanagan LB, Pearcy RW (1991) Climate change and the evolution of C(4) photosynthesis. *Trends Ecol Evol* 6: 95–99. [https://doi.org/10.1016/0169-5347\(91\)90183-X](https://doi.org/10.1016/0169-5347(91)90183-X) PMID: 21232434
5. Liu Z, Sun N, Yang S, Zhao Y, Wang X, Hao X, et al. (2013) Evolutionary transition from C3 to C4 photosynthesis and the route to C4 rice. *Biologia* 68: 577–586.
6. Sage RF, Christin P-A, Edwards EJ (2011) The C4 plant lineages of planet Earth. *Journal of Experimental Botany* 62: 3155–3169. <https://doi.org/10.1093/jxb/err048> PMID: 21414957
7. Simpson MG (2010) Plant Systematics. San Diego, CA, USA: Academic Press.
8. Still CJ, Berry JA, Collatz GJ, DeFries RS (2003) Global distribution of C3 and C4 vegetation: Carbon cycle implications. *Global Biogeochemical Cycles* 17: 6.1–6.14.
9. Gowik U, Westhoff P (2011) The Path from C3 to C4 Photosynthesis. *Plant Physiology* 155: 56–63. <https://doi.org/10.1104/pp.110.165308> PMID: 20940348
10. Edwards EJ, Osborne CP, Strömborg CAE, Smith SA, Consortium CG (2010) The Origins of C4 Grasslands: Integrating Evolutionary and Ecosystem Science. *Science* 328: 587–591. <https://doi.org/10.1126/science.1177216> PMID: 20431008
11. Andrews TJ, and Lorimer G.H. (1987) Rubisco: Structure, mechanisms and prospect for improvement. *The Biochemistry of Plants*: New York: Academic Press. pp. 131–218.
12. Jordan DB, Ogren WL (1981) Species variation in the specificity of ribulose biphosphate carboxylase/oxygenase. *Nature* 291: 513–515.
13. von Caemmerer S, Quick WP (2000) Rubisco: Physiology in Vivo. In: Leegood R, Sharkey T, von Caemmerer S, editors. *Photosynthesis*: Springer Netherlands. pp. 85–113.
14. Monson Russell K. (2003) Gene Duplication, Neofunctionalization, and the Evolution of C4 Photosynthesis. *International Journal of Plant Sciences* 164: S43–S54.
15. Taniguchi Y, Ohkawa H, Masumoto C, Fukuda T, Tamai T, Lee K, et al. (2008) Overproduction of C4 photosynthetic enzymes in transgenic rice plants: an approach to introduce the C4-like photosynthetic pathway into rice. *J Exp Bot* 59: 1799–1809. <https://doi.org/10.1093/jxb/ern016> PMID: 18316317
16. von Caemmerer S, Evans JR (2010) Enhancing C3 Photosynthesis. *Plant Physiology* 154: 589–592. <https://doi.org/10.1104/pp.110.160952> PMID: 20921190

17. Christin PA, Salamin N, Muasya AM, Roalson EH, Russier F, Besnard G (2008) Evolutionary switch and genetic convergence on *rbcL* following the evolution of C4 photosynthesis. *Mol Biol Evol* 25: 2361–2368. <https://doi.org/10.1093/molbev/msn178> PMID: 18695049
18. Iida S, Miyagi A, Aoki S, Ito M, Kadono Y, Kosuge K (2009) Molecular Adaptation of *rbcL* in the Heterophylloous Aquatic Plant Potamogeton. *PloS one* 4: e4633. <https://doi.org/10.1371/journal.pone.0004633> PMID: 19247501
19. Kapralov MV, Filatov DA (2007) Widespread positive selection in the photosynthetic Rubisco enzyme. *BMC evolutionary biology* 7: 73. <https://doi.org/10.1186/1471-2148-7-73> PMID: 17498284
20. Kapralov MV, Smith JAC, Filatov DA (2012) Rubisco Evolution in C(4) Eudicots: An Analysis of Amaranthaceae Sensu Lato. *PloS one* 7: e52974. <https://doi.org/10.1371/journal.pone.0052974> PMID: 23285238
21. Kapralov MV, Votintseva AA, Filatov DA (2013) Molecular Adaptation during a Rapid Adaptive Radiation. *Molecular biology and evolution* 30: 1051–1059. <https://doi.org/10.1093/molbev/mst013> PMID: 23355532
22. Sen L, Fares MA, Liang B, Gao L, Wang B, Wang T, et al. (2011) Molecular evolution of *rbcL* in three gymnosperm families: identifying adaptive and coevolutionary patterns. *Biology Direct* 6: 29–29. <https://doi.org/10.1186/1745-6150-6-29> PMID: 21639885
23. Studer RA, Christin P-A, Williams MA, Orengo CA (2014) Stability-activity tradeoffs constrain the adaptive evolution of RubisCO. *Proceedings of the National Academy of Sciences* 111: 2223–2228.
24. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13: 555–556. PMID: 9367129
25. Murrell B, de Oliveira T, Seebregts C, Kosakovsky Pond SL, Scheffler K, on behalf of the Southern African T, et al. (2012) Modeling HIV-1 Drug Resistance as Episodic Directional Selection. *PLOS Computational Biology* 8: e1002507. <https://doi.org/10.1371/journal.pcbi.1002507> PMID: 22589711
26. Thiltgen G, dos Reis M, Goldstein RA (2017) Finding Direction in the Search for Selection. *Journal of Molecular Evolution* 84: 39–50. <https://doi.org/10.1007/s00239-016-9765-5> PMID: 27913840
27. Tamuri AU, dos Reis M, Goldstein RA (2012) Estimating the Distribution of Selection Coefficients from Phylogenetic Data Using Sitewise Mutation-Selection Models. *Genetics* 190: 1101–1115. <https://doi.org/10.1534/genetics.111.136432> PMID: 22209901
28. Tamuri AU, dos Reis M, Hay AJ, Goldstein RA (2009) Identifying Changes in Selective Constraints: Host Shifts in Influenza. *PLoS Comput Biol* 5: e1000564. <https://doi.org/10.1371/journal.pcbi.1000564> PMID: 19911053
29. Parto S, Lartillot N (2017) Detecting consistent patterns of directional adaptation using differential selection codon models. *BMC evolutionary biology* 17: 147. <https://doi.org/10.1186/s12862-017-0979-y> PMID: 28645318
30. Yang Z, Nielsen R (2002) Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages. *Molecular biology and evolution* 19: 908–917. <https://doi.org/10.1093/oxfordjournals.molbev.a004148> PMID: 12032247
31. Huson DH, Scornavacca C (2012) Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks. *Systematic Biology* 61: 1061–1067. <https://doi.org/10.1093/sysbio/sys062> PMID: 22780991
32. Christin P-A, Freckleton RP, Osborne CP (2010) Can phylogenetics identify C4 origins and reversals? *Trends in Ecology & Evolution* 25: 403–409.
33. Halpern AL, Bruno WJ (1998) Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Molecular biology and evolution* 15: 910–917. <https://doi.org/10.1093/oxfordjournals.molbev.a025995> PMID: 9656490
34. Rodrigue N, Philippe H, Lartillot N (2010) Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proceedings of the National Academy of Sciences of the United States of America* 107: 4629–4634. <https://doi.org/10.1073/pnas.0910915107> PMID: 20176949
35. Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular biology and evolution* 11: 725–736. <https://doi.org/10.1093/oxfordjournals.molbev.a040153> PMID: 7968486
36. Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular biology and evolution* 11: 715–724. <https://doi.org/10.1093/oxfordjournals.molbev.a040152> PMID: 7968485
37. Anisimova M, Bielawski JP, Yang Z (2001) Accuracy and Power of the Likelihood Ratio Test in Detecting Adaptive Molecular Evolution. *Molecular biology and evolution* 18: 1585–1592. <https://doi.org/10.1093/oxfordjournals.molbev.a003945> PMID: 11470850

38. Nielsen R, Yang Z (1998) Likelihood Models for Detecting Positively Selected Amino Acid Sites and Applications to the HIV-1 Envelope Gene. *Genetics* 148: 929–936. PMID: [9539414](#)
39. Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular biology and evolution* 15: 568–573. <https://doi.org/10.1093/oxfordjournals.molbev.a025957> PMID: [9580986](#)
40. Lartillot N (2006) Conjugate Gibbs sampling for Bayesian phylogenetic models. *J Comput Biol* 13: 1701–1722. <https://doi.org/10.1089/cmb.2006.13.1701> PMID: [17238840](#)
41. Lartillot N, Poujol R (2011) A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Molecular biology and evolution* 28: 729–744. <https://doi.org/10.1093/molbev/msq244> PMID: [20926596](#)
42. Lartillot N, Lepage T, Blanquart S (2009) PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25.
43. Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. *Nucleic acids research* 18: 6097–6100. PMID: [2172928](#)
44. van Lun M, van der Spoel D, Andersson I (2011) Subunit Interface Dynamics in Hexadecameric Rubisco. *Journal of Molecular Biology* 411: 1083–1098. <https://doi.org/10.1016/j.jmb.2011.06.052> PMID: [21745478](#)
45. Andersson I, Backlund A (2008) Structure and function of Rubisco. *Plant Physiol Biochem* 46: 275–291. <https://doi.org/10.1016/j.plaphy.2008.01.001> PMID: [18294858](#)
46. Spielman SJ, Wilke CO (2015) The Relationship between dN/dS and Scaled Selection Coefficients. *Molecular biology and evolution*.
47. Tabita FR, Hanson TE, Li H, Satagopan S, Singh J, Chan S (2007) Function, Structure, and Evolution of the RubisCO-Like Proteins and Their RubisCO Homologs. *Microbiology and Molecular Biology Reviews* 71: 576–599. <https://doi.org/10.1128/MMBR.00015-07> PMID: [18063718](#)
48. 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: 609–620. PMID: [11886880](#)