Botanical Studies (2010) 51: 491-510.

Molecular evolution and positive Darwinian selection of the gymnosperm photosynthetic Rubisco enzyme

Da Cheng HAO[1,*], Jun MU[1], and Pei Gen XIAO[2]

[1]*Biotechnology Institute, College of Environment, Dalian Jiaotong University, Dalian 116028, P.R. China*
[2]*Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences, Beijing 100193, P.R. China*

(Received August 27, 2009; Accepted February 9, 2010)

ABSTRACT. Although it was found that gymnosperm *rbc*L of three orders evolves under Darwinian positive selection, it is not clear whether *rbc*L in other gymnosperm lineages is also subject to positive selection. In this study, eleven gymnosperm groups, representing 393 species at various evolutionary levels, were used to illustrate the molecular adaptation and evolutionary dynamics of gene divergence in *rbc*Ls. *rbc*L sequences were amplified from 21 Taxaceae and 10 Cephalotaxaceae species. *rbc*L sequences of other species were retrieved from GenBank. Selective influences were investigated using standard d*N*/d*S* ratio methods and more sensitive techniques investigating the amino acid property changes resulting from nonsynonymous replacements in a phylogenetic context. Analyses revealed the presence of positive selection in *rbc*Ls of all gymnosperm groups. Twenty most often positively selected amino acid sites were characterized. In Taxaceae and Cephalotaxaceae, seven amino acid properties, equilibrium constant of ionization –COOH as the most significant, were found to be influenced by destabilizing positive selection. Some amino acid sites relating to these properties were found to be involved in active site, intradimer interaction, dimer-dimer interaction, and interactions with Rubisco small subunits. Moreover, removing amino acid sites that are under positive selection has significant effect on the bootstrap values of phylogenetic reconstruction. Our results suggest that the conservative *rbc*L evolves under positive selection in gymnosperm lineages. Several regions of *rbc*L have experienced molecular adaptation which fine-tunes photosynthetic Rubisco performance.

Keywords: Chloroplast *rbc*L; Gymnosperm; Maximum likelihood; Physicochemical evolution; Positive selection.

INTRODUCTION

The *rbc*L gene is located in the large single-copy region of the chloroplast genome and encodes large subunits of ribulose-1, 5-bisphosphate carboxylase (Rubisco). The gene is roughly 1425 bp in length, corresponding to 475 amino acids. *rbc*L protein is characterized by a set of eight α helices and eight parallel β strands that ''roll up'' to form a barrel with the β strands on the inside. *rbc*L has been extensively used in determining evolutionary histories at various taxonomic levels (e.g., Müller et al., 2006; Hao et al., 2008), and recently it is recommended to be one of the most useful DNA barcoding markers in species identification (CBOL Plant Working Group, 2009). Moreover, how *rbc*L and small subunits form the functional Rubisco has been studied extensively, with the hope of manipulating this photosynthetic enzyme and increasing crop yield (Christin et al., 2008).

*rbc*L is often chosen for phylogenetic reconstructions and it has been sequenced in thousands of plant species. Surprisingly, despite *rbc*L's physiological importance and abundance of sequence data, *rbc*L is generally used as strings of anonymous nucleotides, without regard to its functional evolution. Kapralov and Filatov (2007) searched for positive selection in *rbc*L sequences from green plants and other phototrophs. Positive selection was found, for the first time, to be present in *rbc*L of most analyzed land plants, but not in algae and cyanobacteria. Positively selected residues are located in regions important for dimer-dimer, intradimer, large subunit-small subunit and Rubisco-Rubisco activase interactions, and that some positively selected residues are close to the active site. Their results demonstrate that despite its conservative nature, *rbc*L gene evolves under positive selection in land plants. Christin et al. (2008) used phylogenetic analyses on a large data set of C_3 and C_4 monocots and found that the *rbc*L gene evolved under positive selection in independent C_4 lineages. This confirms that selective pressures on Rubisco have been switched in C_4 plants by the high CO_2 environment prevailing in their photosynthetic cells. Eight *rbc*L codons evolving under positive selection in C_4 clades were involved in parallel changes among the 23 independent monocot C_4 lineages. The introgression of C_4-like high-efficiency Rubisco would strongly enhance C_3 crop yields in the future CO_2-enriched atmosphere (Christin et al., 2008). Recently, *rbc*L positive selection was also de-

*Corresponding author: E-mail: hao@djtu.edu.cn.

tected among cryptic species in *Conocephalum* (Hepaticae, Bryophytes; Miwa et al., 2009) and the heterophyllous aquatic plant *Potamogeton* (Iida et al., 2009). However, little is known about *rbc*L evolution in gymnosperm. Relatively few gymnosperm species were analyzed for positive selection (Kapralov and Filatov, 2007) and there is no study addressing gymnosperm *rbc*L evolution at the protein level. To gain deeper insight into the evolutionary pattern of *rbc*L of divergent gymnosperm groups, we detect positive selection of *rbc*L in 11 gymnosperm groups at various evolutionary levels. Positive Darwinian selection amino acid site was found in all studied groups, using various likelihood-based methods. We identified positively selected residues in disparate regions of functional importance. The contrasting ecological conditions between gymnosperm and angiosperm as well as among different gymnosperm groups have imposed different selective pressures on Rubisco. The increased amino acid replacement in *rbc*L may reflect the continuous fine-tuning of Rubisco under varying ecological conditions.

MATERIALS AND METHODS

Taxon sampling and data preparation

Sampling of Taxaceae and Cephalotaxaceae species, genomic DNA extraction, PCR amplification of *rbc*L, cloning and DNA sequencing were performed as previously described (Hao et al., 2008). Primers used are: rbcf, 5'-GTCGGATTCAAAGCTGGTGTT-3' and rbcr, 5'-CCTTCATTACGAGCTTGCACA-3', which amplify nearly full-length *rbc*L sequence. Thirty-one *rbc*L sequences were newly generated for this study. Other *rbc*L sequences (mostly full length) used in this study were extracted from NCBI GenBank and the species names and accession numbers as well as taxonomic information are given in Table S1. The obtained sequences were codon-aligned and edited using RevTrans (Wernersson and Pedersen, 2003; http://www.cbs.dtu.dk/services/RevTrans/) and Clustal W2. We analyzed 11 separate data sets (see below). Doubtful sequences (such as containing stop codons) were not included into analyses. All alignments are available upon request from the corresponding author.

Phylogenetic analyses

The best-fit model, JTT, for the amino acid alignment was determined using ProtTest 1.2.6 (Abascal et al., 2005). DNA data were analyzed with Modeltest 3.8 (Posada, 2006) to find the best model of evolution for the data. Employing the Akaike information criterion (AIC), the model with the lowest AIC score was chosen. Neighbor-joining (NJ) analysis was performed by MEGA4 (Tamura et al., 2007). Maximum likelihood (ML) analysis and bootstrapping were performed using RAxML BlackBox (Stamatakis, 2008). The data sets were also analyzed with MrBayes 3.1.2 (Ronquist and Huelsenbeck, 2003). Two independent runs with one cold and three heated Markov chains each per analysis were performed simultaneously until the average standard deviation of split frequencies between the two runs dropped below 0.01. Analyses were run twice to check for consistency of results. We ran two simultaneous runs for 8×10^5 (protein) and 1.8×10^6 (nucleotide) generations, sampling trees every 100 (protein) and 500 (nucleotide) generations, respectively. Topology and branch-length information were summarized in 50% majority rule consensus trees. The *rbc*L sequences of *Podocarpus* were used as the reference for the rooted tree reconstruction.

Molecular evolutionary analysis

Molecular adaptation tests on the *rbc*L codon sites and reconstruction of the ancestral *rbc*L sequences were performed using PAML 4.1 (Yang, 2007). The models used the nonsynonymous/synonymous substitution rate ratio ($\omega = dN/dS$) as an indicator of selective pressure and allowed the ratio to vary among codon sites. We used five site-specific codon substitution models: null models for testing positive selection (M1A, M7, and M8A) and models allowing for positive selection (M2A and M8). The likelihood ratio test (LRT) was used to compare these alternative models. Cases in which M8 model fitted better with $p < 0.05$ in both M7-M8 and M8a-M8 comparisons were regarded as having positive selection.

Because Yang models are based on theoretical assumptions and ignore the empirical observation that distinct amino acids differ in their replacement rates, we also implemented MEC (Mechanistic Empirical Combination) model (Doron-Faigenboim and Pupko, 2007) that takes into account not only the transition-transversion bias and the nonsynonymous/synonymous ratio, but also the different amino acid replacement probabilities as specified in empirical amino acid matrices. Because the LRT is applicable only when two models are nested and thus is not suitable for comparing MEC and M8a models, the second-order AIC (AICc) was used for comparisons (Doron-Faigenboim and Pupko, 2007). Those sites that are most likely to be in the positive selection class ($\omega > 1$) are identified as likely targets of selection.

Recent methods have investigated selection in protein-coding genes further by addressing the type of positive selection detected (directional or nondirectional, stabilizing or destabilizing), the purifying selection, and how the identified selection affects the overall structure and function of the protein. For detecting selection in amino acid sequences we can look at the magnitudes of property change of nonsynonymous residues across a phylogeny. Amino acid substitutions have a wide range of effects on a protein depending on the difference in physicochemical properties and location in the protein structure. This approach provides further resolution to differentiating between types of selective pressures with the ability to detect positive and negative and stabilizing and destabilizing selection and offers insights into the structural and functional consequences of the identified residues under selection (McClellan et al., 2005). We used TreeSAAP v3.2 (Woolley et al., 2003) to test for selection on amino acid

Table S1. Sampling design. List of 11 analyzed groups is provided including taxonomic information and GenBank accession numbers of *rbc*L sequences.

Group No	Group	Order	Family	Genus	Species	GenBank No
1	Gnetales	Gnetales	Gnetaceae	*Gnetum*	*africanum*	AY296526
					cuspidatum	AY296530
					diminutum	AY296532
					gnemon	AY296536
					gnemonoides	AY296539
					hainanense	AY296546
					indicum	AY056574
					klossii	AY296551
					latifolium	AY296553
						AY296555
					macrostachyum	AY296557
					microcarpum	AY296559
						AY296558
					microstachyum	AY296560
					montanum	AY056575
					neglectum	AY296562
					parvifolium	AY056577
						D10734
					schwackeanum	AY296567
					ula	AY296568
					urens	AY296569
					woodsonianum	AY296570
		Welwitschiales	Welwitschiaceae	*Welwitschia*	*mirabilis*	AF394335
2	Ephedrales	Ephedrales	Ephedraceae	*Ephedra*	*alata*	AY755805
					altissima	AY755803
					americana	AY056559
					andina	AY755782
					antisyphilitica	AY755789
					aphylla	AY755802
					californica	AY056569
					chilensis	AY755786
					ciliata	AY755807
					distachya	AY755793
					equisetina	AY056572
					fragilis	AY755784
					frustillata	AY056564
					gerardiana	AY755792
					intermedia	AY056566
					likiangensis	AY755798
					major	AY056571
					minuta	AY755788
					monosperma	AY056561
					nevadensis	AY755796
					procera	AY755795
					przewalskii	EF053223
					rhytidosperma	DQ212957
					rupestris	AY755797
					sinica	D10732
					sp. Tibet-7 isolate 207	EF053225
					Tibet-5	EF053224

Table S1. (Continuation)

Group No	Group	Order	Family	Genus	Species	GenBank No
					torreyana	AY755791
					triandra	EF053227
					trifurca	AY755794
					tweediana	L12677
					viridis	AY056563
3	Coniferales-1	Coniferales	Cupressaceae	*Cupressus*	*abramsiana*	AY988233
					arizonica	AY380886
					austrotibetica	AY988236
					bakeri	AY988237
					benthamii	AY988238
					cashmeriana	AY988240
					chengiana	AY988241
					corneyana	AF479876
					duclouxiana	AY380887
					dupreziana	AY988243
					forbesii	AY988244
					gigantea	AY988246
					glabra	AY988247
					goveniana	AY380888
					guadalupensis	AY988248
					jiangeensis	AY988249
					lusitanica	AY988250
					macnabiana	AY380890
					macrocarpa	AY380891
					montana	AY988252
					nevadensis	AY988253
					pygmaea	AY380892
					sargentii	AY988254
					sempervirens	L12571
					stephensonii	AY988255
					tonkinensis	AY988256
					torulosa	AY988257
				Juniperus	*californica*	AY988258
					coahuilensis	AY988259
					communis	AY988260
					conferta	L12573
					deppeana	AY988261
					drupacea	AY380893
					indica	AY988262
					occidentalis	AY988263
					osteosperma	AY988264
					procera	AY380894
					virginiana	AF119182
						AY988265
				Xanthocyparis	*vietnamensis*	AY380895
4	Coniferales-2	Coniferales	Cupressaceae	*Callitris*	*rhomboidea*	L12537
				Calocedrus	*decurrens*	L12569
					macrolepis	AY380878
				Chamaecyparis	*formosensis*	AY380879
					lawsoniana	AY380880
					obtusa	L12570

Table S1. (Continuation)

Group No	Group	Order	Family	Genus	Species	GenBank No
					pisifera	AY380883
					thyoides	AY380884
				Cunninghamia	*lanceolata*	AY140260
				Cupressus	*sempervirens*	L12571
				Diselma	*archeri*	L12572
				Juniperus	*conferta*	L12573
				Libocedrus	*plumosa*	L12574
				Microbiota	*decussata*	L12575
				Neocallitropsis	*araucarioides*	AF127426
				Platycladus	*orientalis*	L13172
				Tetraclinis	*articulata*	L12576
				Thuja	*occidentalis*	L12578
					plicata	AF127428
						AY237154
				Thujopsis	*dolabrata*	L12577
				Widdringtonia	*cedarbergensis*	L12538
					nodiflora	AY988266
				Sequoiadendron	*giganteum*	AY056580
				Athrotaxis	*laxifolia*	L25754
				Cryptomeria	*japonica*	AJ621937
				Glyptostrobus	*lineatus*	L25750
				Metasequoia	*glyptostroboides*	AJ235805
				Sequoia	*sempervirens*	L25755
				Taiwania	*cryptomerioides*	L25756
				Taxodium	*distichum*	AF119185
5	Coniferales-3	Coniferales	Cephalotaxaceae	*Cephalotaxus*	*harringtonia*	AF227461
					wilsoniana	AB027312
					sinensis	EF660728
					fortunei	AY450863
					latifolia	EF660712
					lanceolata	EF660709
					hainanensis	EF660729
					oliveri	AY450865
					fortunei var. *alpina*	EF660714
					mannii	EF660707
					griffithii	EF660704
					harringtonia var. *drupacea*	EF660716
					koreana	EF660726
					harringtonia cv. *fastigiata*	EF660730
			Taxaceae	*Amentotaxus*	*argotaenia*	EF660731
					formosana	EF660708
					yunnanensis	EF660713
				Austrotaxus	*spicata*	AF456385
				Pseudotaxus	*chienii*	AF456386
				Taxus	*baccata*	AF456388
						EF660721
					brevifolia	AF249666
					mairei	AB027316
						EF660718
					cuspidata	EF660720
					cuspidata var. *nana*	EF660715

Table S1. (Continuation)

Group No	Group	Order	Family	Genus	Species	GenBank No
					yunnanensis	EF660705
					chinensis	EF660719
					×hunnewelliana	EF660723
					wallichiana	EF660717
					fuana (*contorta*)	EF660725
					sumatrana	EF660706
					×media	EF660722
					canadensis	EF660724
					floridana	EF660711
					globosa	EF660710
				Torreya	*yunnanensis*	AY450861
					nucifera	AB027317
					taxifolia	AF456389
					fargesii	EF660735
					californica	EF660732
					grandis	EF660733
					jackii	EF660734
6	Coniferales-4	Coniferales	Pinaceae	*Larix*	*chinensis*	AY389136
					decidua	AB019826
					gmelinii	AY389138
					kaempferi	AB045038
					laricina	AF479878
					occidentalis	X63663
					potaninii	AY389137
				Picea	*bicolor*	AB045041
					chihuahuana	EU269030
					glehnii	AB045042
					jezoensis	AB045043
					maximowiczii	AB045049
					polita	AB045050
						AB045051
					pungens	X58136
					shirasawae	AB045047
					sitchensis	X63660
					smithiana	AF145458
				Pseudotsuga	*menziesii*	X52937
7	Coniferales-5	Coniferales	Pinaceae	*Abies*	*alba*	AB029652
					amabilis	AB029650
					bracteata	AB029647
					firma	AB015647
					hidalgensis	EU269028
					magnifica	AB029649
						X58391
					mariesii	AB015650
					nebrodensis	AB029653
					nordmanniana	AB029654
					numidica	AB029655
					pinsapo	AB029656
					procera	AB029651
				Cedrus	*atlantica*	AF145457
					deodara	AF456381

Table S1. (Continuation)

Group No	Group	Order	Family	Genus	Species	GenBank No
				Keteleeria	*davidiana*	X63664
				Pseudolarix	*amabilis*	DQ987889
					kaempferi	X58782
				Tsuga	*canadensis*	AY056581
					dumosa	AF145460
					mertensiana	AF145463
					forrestii	AF145461
					chinensis	AF145462
					heterophylla	X63659
				Cathaya	*argyrophylla*	AF015786
				Nothotsuga	*longibracteata*	AF145459
8	Coniferales-6	Coniferales	Pinaceae	*Pinus*	*albicaulis*	AY497225
					aristata	AY115758
					attenuata	DQ353724
					ayacahuite	AY497221
					balfouriana	AY115760
					bhutanica	DQ353719
					bungeana	AY115761
					caribaea	AY497244
					catarinae	AY115749
					cembra	DQ353720
					cembroides	AY115751
					cembroides subsp. *lagunae*	AY115752
					cembroides subsp. *orizabensis*	AY115753
					chiapensis	AY497220
					clausa	AY497229
					contorta	AY497230
					cooperi	DQ353723
					coulteri	AY724759
					culminicola	AY115748
					densiflora	DQ353731
					devoniana	AY497241
					discolor	AY115745
					douglasiana	AY497238
					durangensis	AY497240
					echinata	AY724754
					edulis	AY115739
					elliottii	AY724755
					engelmannii	AY497239
					flexilis	AY497222
					gerardiana	AY115762
					glabra	DQ353728
					greggii	AY497246
					hartwegii	AY497231
					heldreichii	DQ353730
					jeffreyi	AY497235
					johannis	AY115747
					juarezensis	AY115743
					kesiya	AY497253
					krempfii	AY115764
					lambertiana	AY497224

Table S1. (Continuation)

Group No	Group	Order	Family	Genus	Species	GenBank No
					leiophylla	AY497243
					longaeva	AY115759
					lumholtzii	AY497242
					massoniana	DQ353732
					maximartinezii	AY115755
					merkusii	AY497251
					monophylla	AY115741
					montezumae	AY497233
					monticola	AY497223
					morrisonicola	AY497227
					muricata	DQ353725
					mugo	EU269032
					nelsonii	AY115757
					nigra	DQ353733
					occidentalis	AY497245
					oocarpa	DQ353726
					palustris	AY724756
					parviflora	EU269033
					patula	AY497248
					peuce	AY497218
					pinceana	AY115754
					pinea	DQ353729
					ponderosa	AY497234
					praetermissa	DQ353727
					pringlei	AY497247
					pseudostrobus	AY497232
					quadrifolia	AY115744
					radiata	AY497250
					remota	AY115750
					resinosa	AY497252
					rigida	AY724757
					roxburghii	AY724760
					rzedowskii	AY115756
					sabiniana	AY497236
					sibirica	AY497228
					serotina	AY724761
					squamata	AY115763
					strobus	AY497219
					taeda	AF119177
					teocote	AY497249
					torreyana	AY497237
					wallichiana	AY734483
					washoensis	DQ353721
9	Coniferales-7	Coniferales	Podocarpaceae	*Phyllocladus*	*trichomanoides*	AB027315
					aspleniifolius	AF249651
					hypophyllus	AF249653
					toatoa	AY442153
				Afrocarpus	*falcatus*	AF249589
					gracilior	X58135
				Dacrycarpus	*imbricatus*	AB027313
					dacrydioides	AF249597

Table S1. (Continuation)

Group No	Group	Order	Family	Genus	Species	GenBank No
					veillardii	AF249598
				Dacrydium	*guillauminii*	AF249635
					araucarioides	AF249632
					balansae	AF249633
					cupressinum	AF249634
				Falcatifolium	*taxoides*	AF249637
				Halocarpus	*kirkii*	AF249640
					bidwillii	AF249638
					biformis	AF249639
				Microstrobos	*niphophilus*	AF249647
					fitzgeraldii	AF249646
				Manoao	*colensoi*	AF249644
				Lepidothamnus	*laxifolius*	AF249643
					fonkii	AF249642
				Sundacarpus	*amarus*	AF249663
				Prumnopitys	*ferruginoides*	AF249659
					andina	AF249655
					ferruginea	AF249656
					ladei	AF249657
					taxifolia	AF249658
				Saxegothaea	*conspicua*	AY664857
				Retrophyllum	*minus*	AF249661
					comptonii	AF249660
				Nageia	*nagi*	AF249648
				Podocarpus	*macrophyllus*	EF660727
						AF249616
					acutifolius	AF249599
					aff. *degeneri*	AF249627
					brassii	AF249601
					chinensis	AF249602
					cunninghamii	AF249603
					dispermus	AF249604
					drouynianus	AF249605
					elatus	AF249606
					gnidioides	AF249607
					grayii	AF249608
					hallii	AF249609
					henkelii	AF249610
					insularis	AF249611
					latifolius	AF249612
					lawrencii	AF249613
					longefoliolatus	AF249614
					lucienii	AF249615
					nivalis	AF249619
					novae-caledoniae	AF249620
					nubigenus	AF307930
					parlatorei	AF249623
					pilgeri	AF249624
					polyspermus	AF249625
					polystachyus	AF249626
					reichei	AF479879

Table S1. (Continuation)

Group No	Group	Order	Family	Genus	Species	GenBank No
					salignus	AF249628
					smithii	AF249629
					spinulosus	AF249630
					totara	AF307931
10	Coniferales-8	Coniferales	Sciadopityaceae	*Sciadopitys*	*verticillata*	L25753
			Araucariaceae	*Agathis*	*borneensis*	AB027310
					australis	AF362993
					dammara	U96477
					lanceolata	U96481
					macrophylla	U87756
					montana	U96478
					moorei	U87755
					obtusa	U96482
					ovata	U87754
					palmerstonii	U96479
					robusta	AF249665
					vitiensis	U96485
				Araucaria	*araucana*	AF249664
					angustifolia	U87750
					bernieri	U96460
					bidwillii	U87751
					columnaris	U96461
					cunninghamii	U87752
					biramulata	U96475
					heterophylla	U96462
					humboldtensis	U96471
					hunsteinii	U87749
					laubenfelsii	U96463
					luxurians	U96464
					meulleri	U87753
					montana	U96457
					nemorosa	U96458
					rulei	U96466
					schmidii	U96473
					scropulorum	U96459
					subulata	U96474
				Wollemia	*nobilis*	AF030419
11	Cycadales	Cycadales	Cycadaceae	*Cycas*	*revoluta*	AY056556
					circinalis	L12674
					micronesica	EU016864
					rumphii	AF394338
					seemannii	AF394340
					thouarsii	AF394336
					wadei	AF394341
				Bowenia	*serrulata*	L12671
					spectabilis	AF531202
		Ginkgoales	Ginkgoaceae	*Ginkgo*	*biloba*	DQ069500

New *rbc*L sequences (EF660704-660735) from this study are in bold type.

properties within our Taxaceae + Cephalotaxaceae data set, for which we can use the species phylogeny resolved in our previous study (Hao et al., 2008). For each property examined, a range of possible 1-step changes as governed by the structure of the genetic code was determined and divided into 8 magnitude categories of equal range, with lower categories indicating more conservative changes and higher categories denoting more radical changes. In order to construct an expected distribution of amino acid property change, each of the 9-nt changes in every codon of every DNA sequence within the data set was evaluated, with each nonsynonymous change assigned to one of the magnitude categories for each property independently. These property changes were then summed across the data set, constructing a set of relative frequencies of change for each of the 8 magnitude categories to establish the null hypothesis under the assumption of neutral conditions (McClellan and McCracken, 2001). If distributions of observed changes fail to fit the expected distributions based on goodness-of-fit scores and z-scores, the null hypothesis of neutrality is rejected. We targeted sites identified to be under positive destabilizing selection, defined as selection for radical amino acid changes resulting in structural or functional shifts in local regions of the protein (McClellan et al., 2005). Positive destabilizing selection is defined as properties with significantly greater amino acid replacements than neutral expectations for magnitude categories 6, 7 and 8 (i.e., the three most radical property change categories). Thirty-one amino acid properties are evaluated across a phylogeny using a sliding window analysis. The results were used to identify regions in the *rbc*L protein that differ significantly from a nearly neutral model at $p = 0.001$. Finally, we identified the particular amino acid residues that contained positive destabilizing selection for each property. These residues might be of general importance to gymnosperm Rubisco function.

Structural analysis of Rubisco

We use published spinach Rubisco protein structure (Taylor et al., 1996; Taylor and Andersson, 1997) for structural analysis. In this study, the numbering of Rubisco large subunit residues is based on the spinach sequence. Rubisco structural data files for spinach 1RBO (Taylor et al., 1996) and 1RCX (Taylor and Andersson, 1997) were obtained from the RCSB Protein Data Bank (http://www.rcsb.org/pdb). The locations and properties of individual amino acids in the Rubisco structure were analyzed using DeepView – Swiss-PdbViewer v.3.7 (Guex and Peitsch, 1997) and confirmed with LPC CSU (Sobolev et al., 1999).

Evaluation of effects of positive selection on phylogenetic reconstructions

Given that positive selection may result in homoplasy we tested whether the removal of codons evolving under positive selection will improve the phylogenetic resolution. We compared boostrap sums of trees reconstructed using all sites (including ones evolving under positive selection) with boostrap sums of trees reconstructed using only neutrally evolving sites. Phylogenetic trees were reconstructed in MEGA4 using NJ algorithm. Gaps were pair-wise deleted. We used 50% majority rule trees and subtracted 50% from each support value before summing up (Kapralov and Filatov, 2007). The subtraction was done to circumvent the bias in summing up bootstrap values of a consensus tree. Without this correction, a tree with two 51% groups would have higher support than one with one group of 100% support, and if support was decreased from 51% to 49%, the sum would be zero (due to a threshold of 50%).

RESULTS AND DISCUSSION

Phylogenetic relationship of gymnosperm *rbc*L proteins

Premature stop codons were not found in all *rbc*L sequences used. Putative amino acid sequences from consensus sequences of cloned *rbc*L genes (21 Taxaceae, 10 Cephalotaxaceae, and one *Podocarpus* taxa) as well as amino acid sequences acquired from GenBank were subjected to a phylogenetic analysis, and a NJ tree generated by MEGA4 is shown in Figure S1. Bayesian analysis and ML method generated the virtually same topology that agrees well with the common view of the conifer topology and is shown in Figure 1. There are two well-supported sister clades, *Cephalotaxus* + *Amentotaxus* + *Taxus* and *Torreya* + *Austrotaxus* + *Pseudotaxus*. *Taxus* (except *T. floridana*) is sister to *Cephalotaxus* + *Amentotaxus*. Within the latter, a *Cephalotaxus* subclade, in which *T. floridana* is included, is basal to other sequences, implying the unique evolutionary pattern of *T. floridana* compared to other *Taxus*. The *rbc*L of *Amentotaxus* is closer to those of *C. koreana*, *C. harringtonia* cv. *fastigiata*, and *C. wilsoniana* than to others. Within *Taxus* there are two sister clades: one consisting of *T. baccata*, *T. cuspidata* and their hybrid, and two North American *Taxus*, the other consisting of *T. canadensis* and Chinese endemic *Taxus*. Within clade *Torreya* + *Austrotaxus* + *Pseudotaxus*, *Austrotaxus* and *Pseudotaxus* are basal to the former. *Torreya jackii* is the first-branching species in *Torreya* clade, while *Torreya californica* is the second one. *Torreya nucifera* is closer to *Torreya taxifolia* than to *Torreya fargesii*. This gene tree is significantly different from both the phylogenetic tree inferred from nuclear ITS and one generated by the combined analysis of five chloroplast DNA markers (Hao et al., 2008). The topology of the *rbc*L tree may reflect, 1. cases where the same amino acid substitution occurred independently in more than one lineage, 2. cases of the retention of plesiomorphic characters, and 3. the possibility of incomplete lineage sorting.

Positive selection in gymnosperm *rbc*L

In order to test for the presence of positive selection acting on *rbc*L we used 403 *rbc*L sequences from 393 gym-

nosperm species (Table S1). These sequences represent six orders and 12 families providing much wider coverage of the gymnosperm lineages than previous study (Kapralov and Filatov, 2007).

For the detection of positive selection we used nested maximum likelihood models allowing for variation in the ratio of non-synonymous to synonymous substitutions rates (d*N*/d*S*) across codons implemented in PAML. We performed two LRTs for the presence of codons under positive selection: M7-M8 and M8a- M8 comparisons. The M7 model assumes a discrete β distribution for d*N*/d*S*, which is constrained between 0 and 1, implemented using ten classes taken in equal proportions. To test for the presence of codons with d*N*/d*S* > 1, M7 is compared to the M8 model, which is similar to the M7 model, but allows for an extra "eleventh" class with d*N*/d*S* ≥ 1. This test was significant for eight out of 11 analyzed groups (Table 1). With the Bonferroni correction (significance level = 0.05/11), this test was significant for seven groups. A more stringent test for positive selection compares model M8 with M8a, which is similar to the model M7, but allows for an extra class of codons with d*N*/d*S* = 1. This test was significant for the same eight groups (Table 1; five groups after the Bonferroni correction). In all cases both M7-M8 and M8a-M8 comparisons rejected models without positive selection in favor of M8 model assuming positive selection (Table 1; five groups after the Bonferroni correction). MEC model (Doron-Faigenboim and Pupko, 2007) takes into account not only the transition–transversion bias and the d*N*/d*S* ratio, but also the different amino acid replacement probabilities as specified in empirical amino acid matrices. Nine out of 11 groups was significant in MEC vs. M8a comparisons (Table 1), except Coniferales-2 (Cupressaceae) and -3 (Taxaceae + Cephalotaxaceae). In these nine groups, model MEC was best-fitting, as the log likelihood value was highest. Compared to M8a, MEC model had much higher log-likelihood value and much lower AICc score in each of these nine groups. Many of the M8 model identified sites (Figure S2) were also identified by MEC model. In two groups in which MEC vs. M8a comparison was negative but the other two types of comparisons were positive, model M8 was best-fitting (Table 1). Thus, we detected *rbc*L positive selection in all

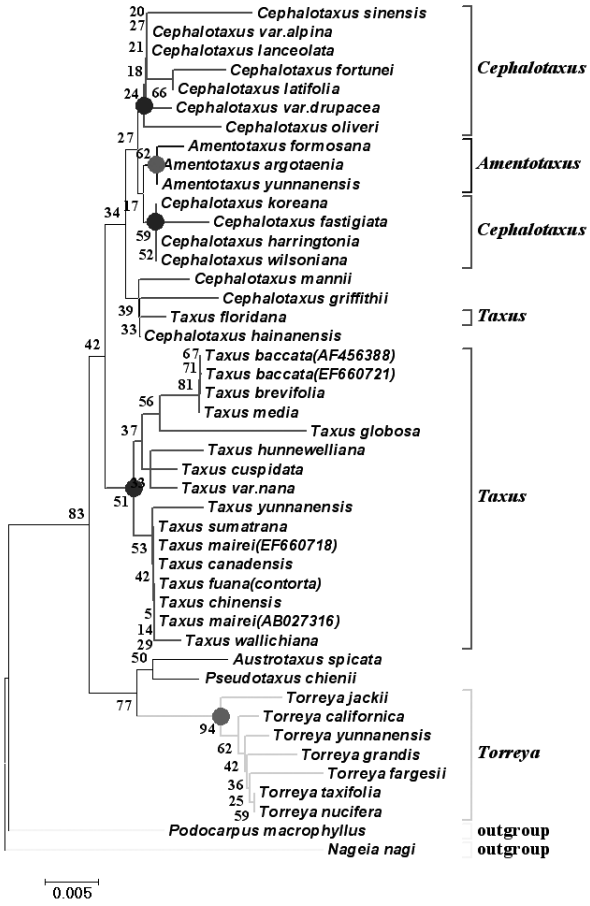

Figure S1. Evolutionary relationships of 45 taxa of Taxaceae, Cephalotaxaceae, and outgroups. The evolutionary history was inferred using the NJ method. The optimal tree with the sum of branch length = 0.202 is shown. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the JTT matrix-based method and are in the units of the number of amino acid substitutions per site. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (Pairwise deletion option). There were a total of 450 positions in the final dataset. Bootstrap values are the percentage of 1000 trials in which a given node was present. Phylogenetic analyses were conducted in MEGA4.

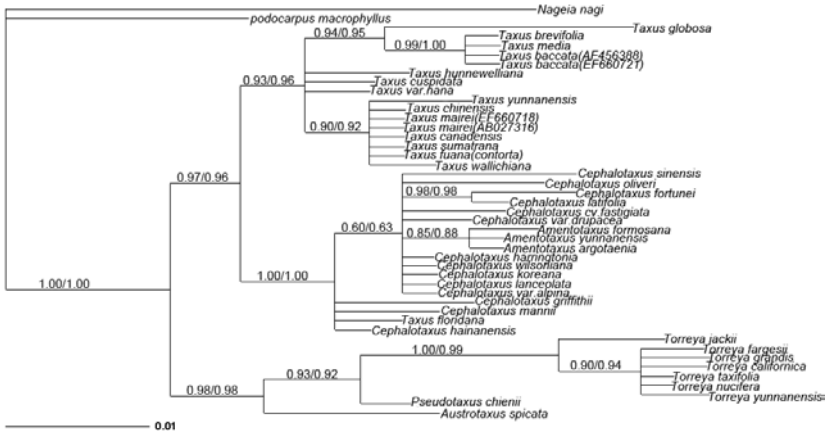

Figure 1. Bayesian 50% majority rule consensus tree (8,000 trees sampled; burn-in = 2,000 trees) inferred from the Taxaceae and Cephalotaxaceae *rbc*L amino acid alignment under the JTT model. Bayesian PPs are given beside branches, before slash (/). ML BPs are given after slash. Branch lengths (scale bar, expected number of substitutions per site) are proportional to the mean of the PPs of the branch lengths of the sampled trees.

Table 1. Likelihood ratio statistics and AICc scores for tests of positive selection.

Plant (group no.)	M8 vs. M8a (df=1)		MEC vs. M8a		M8 vs. M7 (df=2)	
	Log-likelihood	*P*	Log-likelihood	AICc	Log-likelihood	*P*
Gnetales (1)	-2800.77/-2804.03	0.0106	-2777.53/-2804.03	5565.1/5616.0	-2800.77/-2809.83	0.0001
Ephedrales (2)	-2341.63/-2344.22	0.0228	-2327.18/-2344.22	4664.4/4696.4	-2341.63/-2345.71	0.0169
Coniferales-1 (3)	-2280.66/-2278.2	**	-2267.85/-2278.2	4545.7/4564.4	-2280.66/-2280.08	**
Coniferales-2 (4)	-4006.76/-4017.24	0.000005		*	-4006.76/-4025.84	0.0
Coniferales-3 (5)	-3435.58/-3438.76	0.0116		*	-3435.58/-3441.67	0.0022
Coniferales-4 (6)	-2309.17/-2313.86	0.0021	-2288.97/-2313.86	4587.9/4635.7	-2309.17/-2316.45	0.0006
Coniferales-5 (7)	-3172.63/-3184.88	0.000001	-3141.54/-3184.88	6293.1/6377.7	-3172.63/-3192.47	0.0
Coniferales-6 (8)	-3067.57/-3089.06	0.0	-3000.09/-3089.06	6010.2/6186.1	-3067.57/-3103.24	0.0
Coniferales-7 (9)	-8215.36/-8144.36	**	-8098.72/-8144.36	16207.4/16296.7	-8215.36/-8165.62	**
Coniferales-8 (10)	-3391.25/-3399.3	0.00006	-3361.1/-3399.3	6732.2/6806.6	-3391.25/-3403.51	0.000005
Cycadales (11)	-2542.91/-2543.66	**	-2527.15/-2543.66	5064.3/5095.3	-2542.91/-2545.32	**

$-2\Delta\ln L = 2(\ln L_{\text{alternative hypothesis}} - \ln L_{\text{null hypothesis}})$, x^2 distribution.

$AICc = -2\log L + 2p\frac{N}{N-p-1}$, *L*, likelihood, *p*, no. of free parameters, *N*, the sequence length. The smaller the AICc value, the better the model explains the data.

*No positively selected sites found in the protein. **Positive selection in the protein is NON-significant.

gymnosperm groups. It should be noted that, on one hand, there is risk of overestimating the number of positive selection because of multiple-comparison problem, and the correction of significance level might be necessary; on the other hand, using d*N*/d*S* as the sole method by which to detect positive selection is too conservative to detect single adaptive amino acid changes and is thus limited in scope. Interestingly, the highest proportion of cases with detected *matK* positive selection was in gymnosperm (60%), compared to monocot (21%) and other angiosperms (53.5%; Hao et al., 2009). These findings are in accordance with the observation of extensive genomic rearrangement of gymnosperm chloroplast genome (Hirao et al., 2008). Yet notwithstanding, Hirao et al. did not mention the association between rearrangements and positive selection. If positive selection is associated with rearrangements, positive selection of monocots should have higher frequency than that of other angiosperms because more rearrangements occurred in monocots than other angiosperms. Gymnosperms and other plants coexist in many biomes as well as microhabitats, e.g., Gnetales and Podocarpaceae grow with angiosperms. The differential ecological and physiological conditions between gymnosperm and other plants that may have imposed differential selective pressures on Rubisco structure and function need to be further studied. The increased amino acid replacement in *rbc*L may reflect the continuous fine-tuning of Rubisco under varying ecological and physiological conditions.

Positive selection tests at protein level

Selection models that implement d_N/d_S ratios as a criteria for detecting selection are generally not sensitive enough to detect subtle molecular adaptations (McClellan et al., 2005). Actually, detecting positive selection using branch site specific and site prediction methods (like PAML) could often lead to false positive results (Nozawa et al., 2009). It is therefore necessary to employ alternative criteria for the detection of positive selection among sites within generally conservative protein-coding genes. The evolutionary constraints on the slowly evolving *rbc*L would preclude the obvious effects of positive selection by traditional criteria. However, if nonsynonymous substitutions are partitioned by the molecular-phenotypic effects of each, positive selection for radical amino acid changes that may have a slower rate but occur more frequently than expected by chance may be more easily detected.

Significant physicochemical amino acid changes among residues in Taxaceae + Cephalotaxaceae *rbc*L were identified by TreeSAAP, which compares the observed distribution of physicochemical changes inferred from a phylogenetic tree with an expected distribution based on the assumption of completely random amino acid replacement expected under the condition of selective neutrality. There are radical *pK'* (equilibrium constant of ionization –COOH, a chemical amino acid property) changes on 44 sites (z score > 3.09, p < 0.001, category 8), among which 20 sites have z score >4 (Table 3). Interestingly, sites 225, 226, and 255 that are involved in interaction with the small subunit and dimer-dimer interaction were also detected by ML-based models (Table 2). Nine sites (sites 22, 180-187) were under category 8 positive destabilizing selection in two chemical amino acid properties, i.e., E_l (long-range non-bonded energy) and H_p (surrounding hydrophobicity). These sites are involved in intradimer interaction, dimer-dimer interaction, and interaction with the small subunit. In addition, sites under category 7 or 6 positive selection

in E_{sm} (short and medium-range non-bonded energy), M_w (molecular weight), H (hydropathy), and V^0 (partial specific volume) are summarized in Table 3. Totally there are five chemical property changes, compared to only one structural property change and one other category change, during the last 66 myr since the origin of Taxaceae and Cephalotaxaceae (Hao et al., 2008). In contrast, P_c (coil tendencies), C_a (helical contact area), P_t (turn tendencies), and α_c (power to be at the C-terminal) were found to undergo category 8 negative (purifying) destabilizing selection. These chemical and conformational amino acid properties (Gromiha and Ponnuswamy, 1993) may well be important to the overall optimization of *rbc*L function in gymnosperm and have been periodically adjusted during cladogenesis to maximize the biochemical effect of the spatial relationships between α-helices/β-sheets/loops and the primary functional amino acid residues that influence the catalytic function of Rubisco.

Distribution of *rbc*L residues that are responsible for the positive selection

The average number of amino acids under selection

Table 2. Twenty most often positively selected rbcL residues in gymnosperm.

Residue No[1]	N[2] (group no.)	Location of residue	Residues within 5 Å[3]	Structural motifs within 5 Å	Interactions[4]
225	7 (1, 3-7, 9)	Helix 2	189, 190, 193, 194, 221, 222, 223, 224, **226**, 227, 228, 229, 236, 237, 238	Helixes 1, 2; strand 3	SSU
95	6 (1, 4-6, 8, 9)		42, 43, 44, 93, 94, 96, 97, 131	Strands B, D, E	ID, RA
449	6 (1, 2, 7-10)	Helix G	445, 446, 447, 448, 450, 451, 452, 453, 455, 456	C-terminus	SSU
375	5 (1, 3, 4, 8, 9)	Strand 7	155, 158, 159, 169, 324, 325, 326, 373, 374, 376, 377, 397, 398, 399	Helix E; strands 6, 7, 8	SSU
255	4 (6-8, 10)	Helix 3	**251**, 252, 253, **254**, 256, 257, 258, 259, 283	Helixes 3, 4	SSU, DD
30	3 (4, 5, 9)		26, 27, **28**, 29, **31**, 32, 85	Strands A, C	ND
31	3 (4, 5, 9)		29, **30**, 32, 33, 35, 37, 85, 102, 139	Strands B, C, D, E	ND
226	3 (1, 7, 8)	Helix 2	221, 222, 223, 224, **225**, 227, 228, 229, 230, 260, 261, 262	Helixes 2, 3	SSU
251	3 (3, 4, 8)	Helix 3	247, 248, 249, 250, 252, 253, **254**, **255**, 256, **279**, 280, 283	Helixes 3, 4	DD, SSU
474	3 (4, 7, 8)		305, 338, 341, 471, 472, 473, 475	Helix 6	ND
28	2 (5, 6)	N-terminus	25, 26, 27, 29, **30**, 84	Strands A, C	ND
50	2 (4, 8)	Helix B	44, 48, 49, 51, 52, 53, 54, 55, 87, 97, 99	Strands B, C, D, helix B	ID
142	2 (1, 2)	Helix D	33, 140, 141, **143**, 144, 145, 146, 367, 369	N-terminus; strands D, H	DD
143	2 (1, 8)	Helix D	34, 105, 141, **142**, 144, 145, 146, 147	Helix D	DD
219	2 (4, 10)	Helix 2	58_I, 59_I, 61_I, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 256, 260	Helixes 2, 3	SSU, DD
254	2 (8, 10)	Helix 3	242, 250, **251**, 252, 253, **255**, 256, 257, 258, 265, 280, 283, 289	Loop 3; helix 3, 4; strand 4	DD, SSU
279	2 (1, 7)	Helix 4	250, **251**, 274, 275, 276, 277, 278, 280, 281, 282, 283	Helixes 3, 4	ND
328	2 (6, 11)	Loop 6	295, 311, 326, 327, 329, 330, 342, 345, 346, 349, 376, 377, 378, 394	AS; loop 6; helixes 5, 7; strand 7	AS
434	2 (1, 2)		430, 431, 432, 433, 435, 436	Helix 8	SSU
466	2 (2, 10)		386, 463, 464, 465, 467, 468		ID

[1]Numbering of residues is after the spinach Rubisco sequence.

[2]Number of groups with detected signal of positive selection where the particular residue was shown under positive selection with Bayesian posterior probability larger than 0.95, when analyzed by the Bayes Empirical Bayes of PAML.

Group 1, Gnetales; 2, Ephedrales; 3, Coniferales-1 (Cupressaceae, *Cupressus* + *Juniperus*); 4, Coniferales-2 (Cupressaceae); 5, Coniferales-3 (Taxaceae + Cephalotaxaceae); 6, Coniferales-4 (Pinaceae, *Larix* + *Picea*); 7, Coniferales-5 (Pinaceae, *Abies* + *Tsuga*); 8, Coniferales-6 (Pinaceae, *Pinus*); 9, Coniferales-7 (Podocarpaceae); 10, Coniferales-8 (Araucariaceae); 11, Cycadales.

[3]Subscriptions denote residues from I small subunit. Residues within the list of the twenty designated residues are boxed.

[4]Interactions in which the twenty selected residues and/or residues within 5 Å of them are involved. AS, interactions with the active site; ID, intradimer interactions; DD, dimer-dimer interactions; RA, interface for interactions with Rubisco activase; SSU, interactions with small subunits; ND, not determined.

per group was 9.6 ± 5.1, e.g., the beta and ω model estimated that only 1.81% of *rbc*L sites of Taxaceae + Cephalotaxaceae have experienced strong positive selection (ω_s = 3.06). In 11 groups with positive selection detected by M7-M8, M8a-M8 or M8a-MEC comparisons, 63 out of 476 Rubisco residues (Table S2) were found to be under positive selection. In all groups more than one residue was under selection. The distribution of residues identified was highly uneven: twenty most often selected residues are responsible for 59.4% of the cases of positive selection (Figure S2, Tables 2 and S2). Analyses of Rubisco tertiary structure revealed that some of the 20 most often selected residues are quite close to each other and most of them are involved in interactions between Rubisco large and small subunits, in interactions with Rubisco activase, dimer-dimer and intradimer interactions, as well as in interactions with the active site (Figure 2, Tables 2 and 3). The analyses of mutant Rubisco enzymes have shown that interface between large and small subunits contributes to holoenzyme thermal stability, catalytic efficiency, and CO_2/O_2 specificity (Spreitzer et al., 2005; Karkehabadi et al., 2005). Rubisco activase is involved in the opening of the closed Rubisco form to release ribulose-1,5-bisphosphate and to produce the active enzyme (Ott et al., 2000; Spreitzer and Salvucci, 2002). Pedron et al. (2009) found that among cold-responsive genes, the expression of Rubisco activase was down-regulated by the 3°C treatment in five cypress genotypes. On the other hand, there was evidence for the adaptive evolution of *rbc*L during diversification in temperature tolerance of hot spring cyanobacteria (Miller, 2003). Way and Sage (2008) found that black spruce HT seedlings (grow at 30/22°C day/night temperatures) at 40°C might be limited by Rubisco capac-

Figure S2. Positively and negatively selected amino acid sites in Taxaceae + Cephalotaxaceae (group 5) detected by M8 model. Scale 1 represents the strongest positive selection and scale 7 represents the strongest negative (purifying) selection. Site 1 corresponds to amino acid site 11 of spinach *rbc*L protein.

Figure 2. Locations of the twenty most often positively selected Rubisco residues. The large subunit of spinach Rubisco is shown (chain L) with locations of the twenty most often positively selected Rubisco residues (Table 2) highlighted by pink circles. Visualization is made using the Cn3D viewer (http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml).

Table 3. Amino acid properties under positive destabilizing selection identified from 31 amino acid properties in TreeSAAP.

Amino acid property		Category, z score (p value)	Site[1]	Function[3]
C	*pK'* (equilibrium constant of ionization –COOH)	8, 5.239 (< 0.001)	220-**225**, **226**-229, **255**, 306, 307, 310-313, 325, 326, 340[2]	SSU, DD
C	E_l (long-range non-bonded energy)	8, 2.121 (< 0.05)	22, 180-187	SSU, DD
C	H_p (surrounding hydrophobicity)	8, 2.121 (< 0.05)	22, 180-187	ID
C	E_{sm} (short and medium-range non-bonded energy)	7, 1.858 (< 0.05)	422-431, 441, 442	SSU
O	M_w (molecular weight)	6, 4.177 (< 0.001)	425-431, 446-**449**, 450-452	SSU
C	*H* (hydropathy)	6, 2.842 (< 0.01)	185-187, 388-397	SSU
S	V^0 (partial specific volume)	6, 2.137 (< 0.05)	425, 433, 441-**449**, 450-452	SSU

[1]Numbering of residues is after the spinach Rubisco sequence. Residues under positive selection, detected by ML-based models, are boxed.

[2]Sites with z score > 4.0 in sliding window analysis are shown.

[3]ID, intradimer interactions; DD, dimer-dimer interactions; SSU, interactions with small subunits; C, chemical; S, structural; O, other.

Table S2. RbcL residues under positive selection.

Residue No	Structural motif	N groups where selected	Analyzed Group No										
			1	2	3	4	5	6	7	8	9	10	11
11		1									+		
13		1									+		
14		1									+		
19		1	+										
28	N-terminus	2					+	+					
30		3				+	+				+		
31		3				+	+				+		
32		1									+		
33		1		+									
45		1	+										
50		2				+				+			
53		1				+							
55		1		+									
86	Strand C	1										+	
87	Strand C	1	+										
89	Strand C	1						+					
91		1						+					
94		1	+										
95		6	+			+	+	+		+	+		
97	Strand D	1									+		
99		1									+		
100		1						+					
116		1				+							
133	Strand E	1									+		
142	Helix D	2	+	+									
143	Helix D	2	+							+			
170		1					+						
178		1										+	
208		1									+		
216		1									+		
219	Helix 2	2				+						+	
225	Helix 2	7	+		+	+	+	+	+		+		
226	Helix 2	3	+						+	+			
251	Helix 3	3			+	+				+			
254	Helix 3	2								+		+	
255	Helix 3	4						+	+	+		+	
265	Strand 4	1							+				
279	Helix 4	2	+						+				
305		1						+					
306		1			+								
320	Helix 5	1	+										
326	Strand 6	1	+										
328	Loop 6	2						+					+
340	Helix 6	1						+					
341	Helix 6	1						+					
365		1		+									
371		1	+										
375	Strand 7	5	+		+	+				+	+		
392		1	+										
418	Helix 8	1								+			

Table S2. (Continuation)

Residue No	Structural motif	N groups where selected	Analyzed Group No										
			1	2	3	4	5	6	7	8	9	10	11
427	Helix 8	1					+						
428	Helix 8	1	+										
434		2	+	+									
449	Helix G	6	+	+					+	+	+	+	
450		1									+		
451		1									+		
452		1	+										
461		1							+				
462		1		+									
466		2		+								+	
471		1								+			
472		1											+
474		3				+			+	+			
Total		106	19	8	4	11	7	11	8	12	17	7	2

Group 1, Gnetales; 2, Ephedrales; 3, Coniferales-1 (Cupressaceae, *Cupressus* + *Juniperus*); 4, Coniferales-2 (Cupressaceae); 5, Coniferales-3 (Taxaceae + Cephalotaxaceae); 6, Coniferales-4 (Pinaceae, *Larix* + *Picea*); 7, Coniferales-5 (Pinaceae, *Abies* + *Tsuga*); 8, Coniferales-6 (Pinaceae, *Pinus*); 9, Coniferales-7 (Podocarpaceae); 10, Coniferales-8 (Araucariaceae); 11, Cycadales.

ity and acclimation, but not by heat lability of Rubisco activase. Moreover, C_3 plants exhibit different Rubisco catalytic properties following the mean temperature that they encounter, i.e., C_3 plants from cooler habitats having a Rubisco with a higher turn-over rate, like C_4 plants (Sage, 2002). Taken together, Rubisco has evolved to improve performance in the environment that plants normally experience. There could be positive selection of gymnosperm (belonging to C_3 plant) *rbc*L in response to various thermal conditions in the respective ecological niche.

Other selective pressures could have driven Rubisco molecular evolution in gymnosperm. For example, specificity factors of Rubisco of C_3 plants vary according to the environmental xericity, i.e., C_3 plants from more arid habitats having a Rubisco with a higher CO_2 specificity (Galmés et al., 2005). Detection of positive selection at the interfaces between chloroplast- and nuclear-encoded Rubisco subunits and between Rubisco and Rubisco activase suggests that co-evolution of proteins in the Rubisco complex can be another driving force of adaptive evolution in *rbc*L.

We found site 225 of helix 2 is the most often positively selected *rbc*L residue in gymnosperm, while it is also one of the most often positively selected residues in angiosperm (Kapralov and Filatov, 2007), although the exact reason is unknown. Loop 6 plays a major role in discrimi-

Table S3. Impact of sites evolving under positive selection on phylogenetic resolution: NJ method.

Group	No. of sequences	Bootstrap sum of NJ 50% majrule tree		Δ, %
		All codons	Without codons evolving under positive selection	
Gnetales + Welwitschiales	24	16	67	318.8
Ephedrales	32	47	114	142.6
Coniferales-1	40	20	28	40
Coniferales-2	31	54	21	-61.1
Coniferales-3 (no outgroup)	43	233	192	-17.6
Coniferales-3 (with outgroup)	45	243	204	-16.0
Coniferales-4	19	269	124	-53.9
Coniferales-5	26	238	196	-17.6
Coniferales-6	83	317	200	-36.9
Coniferales-7	63	119	183	53.8
Coniferales-8	33	282	267	-5.3
Cycadales+Ginkgoales	10	160	243	51.9

nating between CO_2 and O_2 and functions as a flexible "flap" that closes over the active site once the substrates are bound (Satagopan and Spreitzer, 2004). In the present study, site 328 of loop 6 was found to be under positive selection in Coniferales-4 (Pinaceae, *Larix* + *Picea*) and Cycadales, which is less common than in angiosperm (Kapralov and Filatov, 2007). Mutation of site 328, by affecting the movement of loop 6, could alter the interaction with the six-carbon intermediates and thus change the CO_2/O_2 specificity of the Rubisco (Christin et al., 2008). More specifically, the effects of amino acid replacements in residue 449 were tested by directed mutagenesis in the green alga *Chlamydomonas reinhardtii*: cystein 449 to serine substitution showed an increased resistance to inactivation when Rubisco in the oxidized state (Marin-Navarro and Moreno, 2006). It is suggested that amino acids evolving under positive selection in *rbc*L are located in regions important for Rubisco activity and residues involved in dimer-dimer, intradimer, large subunit-small subunit and Rubisco-Rubisco activase interactions as well as ones close to the active site are the prime targets of positive selection in Rubisco (Table 2). It is apparent that gymnosperm Rubisco share an interesting history and undoubtedly present a classic example of divergent evolution. The different gymnosperm Rubiscos found in nature, some of which must function in extreme or inhospitable environments, have made structural adaptations to allow catalysis to occur. Effects of the positively selected sites have to be depicted through structural analyses and these sites should be mutated, both alone and in combination. The Rubisco regions characterized by high density of residues evolving under positive selection and located relatively far away from the active site could be good candidates for mutagenic studies to reveal the broader picture of how gymnosperm Rubisco functions.

Implications for phylogenetic studies

Our analysis demonstrated that *rbc*L can not be regarded as a neutral marker and positive selection is not unusual in gymnosperm. Positive selection may result in homoplasy due to fixations of the same mutation that arose independently in several phylogenetic lineages (Figure 1 and S1). We tested whether the removal of codons evolving under positive selection will improve phylogenetic resolution (Table S3). We compared sums of bootstrap values between the trees reconstructed using all sites and the trees reconstructed using only neutrally evolving sites (positively selected sites excluded). The sums of bootstrap frequencies decreased for more than 5% in six groups, and increased for more than 5% in five groups (Gnetales 318.8%, Ephedrales 142.6%, Coniferales-1, 40%, Coniferales-7, 53.8%, and Cycadales 51.9%). The putative positive selection sites or putatively potential homoplasy characters are not necessarily parsimony informative sites and, hence, may leave no effect on the reconstructed MP topology. Yet, we compared sums of bootstrap values between the MP trees reconstructed using all sites and those reconstructed using only neutrally evolving sites and found similar results (data not shown). Thus, taking into account the presence of positive selection in *rbc*L may improve phylogenetic reconstructions in the specific groups. *rbc*L datasets should be checked for positive selection, and if selection is found, whether deletion of sites evolving under positive selection would increase topological resolution/bootstrap support should be tested. Previously we found strong cytonuclear incongruence partially caused by positive selection in *mat*K and *rbc*L in Taxaceae (Hao et al., 2008, 2009). This exemplifies the risk of reconstructing phylogenetic and phylogenomic relations solely from chloroplast data in groups with interspecific hybridization. Tests for the presence of positive selection and for the congruence between chloroplast and nuclear phylogenies are indispensable for correct inference of species phylogenetic and phylogenomic relationships.

Acknowledgements. This study is supported by the Education Department of Liaoning Province (2009A120), and Start-up research fund (2008-2010) of Dalian Jiaotong University. The authors are grateful to Ms. Yutian Liang (Dalian Jiaotong University) for her help in structural analysis of Rubisco.

LITERATURE CITED

CBOL Plant Working Group. 2009. A DNA barcode for land plants. Proc. Natl. Acad. Sci. USA **106:** 12794-12797.

Christin, P.A., N. Salamin, A.M. Muasya, E.H. Roalson, F. Russier, and G. Besnard. 2008. Evolutionary switch and genetic convergence on *rbc*L following the evolution of C4 photosynthesis. Mol. Biol. Evol. **25:** 2361-2368.

Doron-Faigenboim, A. and T.A. Pupko. 2007. Combined empirical and mechanistic codon model. Mol. Biol. Evol. **24:** 388-397.

Galmés, J., J. Flexas, A.J. Keys, J. Cifre, R.A.C. Mitchell, P.J. Madgwick, R.P. Haslam, H. Medrano, and M.A.J. Parry. 2005. Rubisco specificity factor tends to be larger in plant species from drier habitats and in species with persistent leaves. Plant Cell Environ. **28:** 571-579.

Gromiha, M.M. and P.K. Ponnuswamy. 1993. Relationship between amino acid properties and protein compressibility. J. Theo. Biol. **165:** 87-100.

Guex, N. and M.C. Peitsch. 1997. SWISS-MODEL and the Swiss-Pdb-Viewer: An environment for comparative protein modeling. Electrophoresis **18:** 2714-2723.

Hao, D.C., S.L. Chen, and P.G. Xiao. 2009. Molecular evolution and positive Darwinian selection of the chloroplast maturase matK. J. Plant Res. DOI 10.1007/s10265-009-0261-5.

Hao, D.C., P.G. Xiao, B. Huang, G.B. Ge, and L. Yang. 2008. Interspecific relationships and origins of Taxaceae and Cephalotaxaceae revealed by partitioned Bayesian analyses of chloroplast and nuclear DNA sequences. Plant Syst. Evol. **276:** 89-104.

Hirao, T., A. Watanabe, M. Kurita, T. Kondo, and K. Takata. 2008. Complete nucleotide sequence of the *Cryptomeria*

japonica D. Don. chloroplast genome and comparative chloroplast genomics: diversified genomic structure of coniferous species. BMC Plant Biol. **8:** 70.

Iida, S., A. Miyagi, S. Aoki, M. Ito, Y. Kadono, and K. Kosuge. 2009. Molecular adaptation of *rbc*L in the heterophyllous aquatic plant *Potamogeton*. PLoS ONE **4:** e4633.

Kapralov, M.V. and D.A. Filatov. 2007. Widespread positive selection in the photosynthetic Rubisco enzyme. BMC Evol. Biol. **7:** 73.

Karkehabadi, S., T.C. Taylor, R.J. Spreitzer, and I. Andersson. 2005. Altered intersubunit interactions in crystal structures of catalytically compromised ribulose-1,5-bisphosphate carboxylase/oxygenase. Biochemistry **44:** 113-120.

Marin-Navarro, J. and J. Moreno. 2006. Cysteines 449 and 459 modulate the reduction-oxidation conformational changes of ribulose 1,5-bisphosphate carboxylase/oxygenase and the translocation of the enzyme to membranes during stress. Plant Cell Environ. **29:** 898-908.

McClellan, D.A. and K.G. McCracken. 2001. Estimating the influence of selection on the variable amino acid sites of the cytochrome b protein functional domains. Mol. Biol. Evol. **18:** 917-925.

McClellan, D.A., E.J. Palfreyman, M.J. Smith, J.L. Moss, R.G. Christensen, and J.K. Sailsbery. 2005. Physiocochemical evolution and molecular adaptation of the cetacean and artiodactyl cytochrome b proteins. Mol. Biol. Evol. **22:** 437-455.

Miller, S.R. 2003. Evidence for the adaptive evolution of the carbon fixation gene *rbc*L during diversification in temperature tolerance of a clade of hot spring cyanobacteria. Mol. Ecol. **12:** 1237-1246.

Miwa, H., I. J. Odrzykoski, A. Matsui, M. Hasegawa, H. Akiyama, Y. Jia, R. Sabirov, H. Takahashi, D. E. Boufford, and N. Murakami. 2009. Adaptive evolution of *rbc*L in *Conocephalum* (Hepaticae, bryophytes). Gene **441:** 169-175.

Müller, K.F., T. Borsch, and K.W. Hilu. 2006. Phylogenetic utility of rapidly evolving DNA at high taxonomical levels: contrasting *mat*K, *trnT-F* and *rbc*L in basal angiosperms. Mol. Phylogenet. Evol. **41:** 99-117.

Nozawa, M., Y. Suzuki, and M. Nei. 2009. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. Proc. Natl. Acad. Sci. USA **106:** 6700-6705.

Ott, C.M., B.D. Smith, A.R. Jr. Portis, and R.J. Spreitzer. 2000. Activase region on chloroplast Ribulose-1,5-bisphosphate carboxylase/oxygenase. J. Biol. Chem. **275:** 26241-26244.

Pedron, L., P. Baldi, A.M. Hietala, and N. La Porta. 2009. Genotype-specific regulation of cold-responsive genes in cypress (*Cupressus sempervirens* L.). Gene **437:** 45-53.

Sage, R.F. 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.

Satagopan, S. and R.J. Spreitzer. 2004. Substitutions at the Asp-473 latch residue of *Chlamydomonas* ribulose-bisphosphate carboxylase/oxygenase cause decreases in carboxylation efficiency and CO_2/O_2 specificity. J. Biol. Chem. **279:** 14240-14244.

Sobolev, V., A. Sorokine, J. Prilusky, E.E. Abola, and M. Edelman. 1999. Edelman. Automated analysis of interatomic contacts in proteins. Bioinformatics **15:** 327-332.

Spreitzer, R.J., S.R. Peddi, and S. Satagopan. 2005. Phylogenetic engineering at an interface between large and small subunits imparts landplant kinetic properties to algal Rubisco. Proc. Natl. Acad. Sci. USA **102:** 17225-17230.

Spreitzer, R.J. and M.E. Salvucci. 2002. RUBISCO: structure, regulatory interactions, and possibilities for a better enzyme. Annu. Rev. Plant Biol. **53:** 449-475.

Tamura, K., J. Dudley, M. Nei, and S. Kumar. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. **24:** 1596-1599.

Taylor, T.C. and I. Andersson. 1997. The structure of the complex between rubisco and its natural substrate ribulose 1,5-bisphosphate. J. Mol. Biol. **265:** 432-444.

Taylor, T.C., M. D. Fothergill, and I. Andersson. 1996. A common structural basis for the inhibition of ribulose 1,5-bisphosphate carboxylase by 4-carboxyarabinitol 1,5-bisphosphate and xylulose 1,5-bisphosphate. J. Biol. Chem. **271:** 32894-32899.

Way, D.A. and R.F. Sage. 2008. Thermal acclimation of photosynthesis in black spruce [*Picea mariana* (Mill.) B.S.P.]. Plant Cell Environ. **31:** 1250-1262.

Wernersson, R. and A.G. Pedersen. 2003. RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. Nucl. Acid. Res. **31:** 3537-3539.

Woolley, S., J. Johnson, M.J. Smith, K.A. Crandall, and D.A. McClellan. 2003. TreeSAAP: selection on amino acid properties using phylogenetic trees. Bioinformatics **19:** 671-672.

Yang, Z. 2007. PAML 4: a program package for phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24:** 1586-1591.

裸子植物光合酶 Rubisco 的分子進化和正達爾文選擇

郝大程[1] 穆 軍[1] 肖培根[2]

[1] 中國大連交通大學 環境學院生物技術研究所

[2] 北京中國醫學科學院 藥用植物研究所

已發現有三目裸子植物 *rbc*L 的進化中經受正選擇作用，但在其他譜系的情形如何尚屬未知。本研究首次全面挖掘 11 個裸子植物類群，393 種植物的 *rbc*L 序列，分析基因歧異過程中的分子適應和進化動力學。PCR 擴增 21 種紅豆杉和 10 種三尖杉的 *rbc*L 序列。從 GenBank 獲取其他種的 *rbc*L 序列。除了標準的 dN/dS 比值法，還將系統發育資訊與非同義替換引起的氨基酸物理化學性質變化聯繫起來，提高了檢測正選擇位點的靈敏性。發現所有裸子植物類群的 *rbc*Ls 的進化均受正選擇作用。重點研究了 20 個最常見的正選擇位點的性質。發現在紅豆杉科和三尖杉科，有 7 個氨基酸性質受到正不穩定選擇作用，其中以羧基端電離平衡常數最顯著。發現與這些物化性質有關的一些氨基酸位點與酶活性位點，二聚體內互作，二聚體間互作，及與 Rubisco 小亞單位的互作均有關聯。移除正選擇氨基酸位元點對系統發育重建的 bootstrap 值有顯著影響。本研究提示進化上保守的裸子植物 *rbc*L 確實經受正選擇作用。*rbc*L 蛋白的不同區域均經歷分子適應以便精確調整酶蛋白功效。

關鍵詞：葉綠體 *rbc*L；最大似然法；正選擇；裸子植物；物理化學進化。