Current Biology 17, 1241–1247, July 17, 2007 ©2007 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2007.06.036

Report

C_4 Photosynthesis Evolved in Grasses via Parallel Adaptive Genetic Changes

Pascal-Antoine Christin,[1,*] Nicolas Salamin,[1] Vincent Savolainen,[2] Melvin R. Duvall,[3] and Guillaume Besnard[1,*]
[1] Department of Ecology and Evolution
Biophore
University of Lausanne
1015 Lausanne
Switzerland
[2] Jodrell Laboratory
Royal Botanic Gardens, Kew
TW9 3DS Surrey
United Kingdom
[3] Department of Biological Sciences
Northern Illinois University
DeKalb, Illinois 60115

Summary

Phenotypic convergence is a widespread and well-recognized evolutionary phenomenon. However, the responsible molecular mechanisms remain often unknown mainly because the genes involved are not identified. A well-known example of physiological convergence is the C_4 photosynthetic pathway, which evolved independently more than 45 times [1]. Here, we address the question of the molecular bases of the C_4 convergent phenotypes in grasses (Poaceae) by reconstructing the evolutionary history of genes encoding a C_4 key enzyme, the phospho*enol*pyruvate carboxylase (PEPC). PEPC genes belong to a multigene family encoding distinct isoforms of which only one is involved in C_4 photosynthesis [2]. By using phylogenetic analyses, we showed that grass C_4 PEPCs appeared at least eight times independently from the same non-C_4 PEPC. Twenty-one amino acids evolved under positive selection and converged to similar or identical amino acids in most of the grass C_4 PEPC lineages. This is the first record of such a high level of molecular convergent evolution, illustrating the repeatability of evolution. These amino acids were responsible for a strong phylogenetic bias grouping all C_4 PEPCs together. The C_4-specific amino acids detected must be essential for C_4 PEPC enzymatic characteristics, and their identification opens new avenues for the engineering of the C_4 pathway in crops.

Results and Discussion

Congruence between Gene and Species Trees Recovered Only with Nearly Neutral Sites

We constituted a data set of 169 PEPC encoding genes, of which 127 were sequenced in this study. In the phylogenetic tree inferred on PEPC coding sequences (Figure 1), all grass genes encoding C_4 PEPC (*ppc-C4*), except that of *Centropodia*, cluster together whereas its closest non-C_4 genes (referred to as *ppc-B2*) form a paraphyletic group. The same pattern occurred whether based on amino acid or nucleotide sequences and regardless of the phylogenetic method used. The species relationships deduced from *ppc-C4* as well as from *ppc-B2* were highly incongruent with the species tree inferred by other markers [3–5]. The obtained gene tree can be explained only by postulating a very high number of gene duplications and losses or horizontal gene transfers, making this topology very unlikely. Because the C_4 trait evolved several times independently in the grass family [1, 3], a single origin of the C_4 PEPC was unexpected.

A potential source of bias, which could be responsible for the *ppc-C4* grouping found in our analyses, is the evolutionary forces driving C_4 PEPC evolution. Because of the crucial role played by this PEPC isoform in the C_4 photosynthetic pathway [1], it could have been the target of strong selective pressures that would have drastically altered the amino acid sequences. If a high enough number of identical amino acids appeared independently in the different C_4 lineages, the codon positions that determine the transition between non-C_4 and C_4-characteristic amino acids would tend to group the *ppc-C4* together and thus be misleading in a phylogenetic context. This hypothesis would then predict that the trees constructed with sites less affected by selection, for example the third positions of the codons or the intron sequences, would have a different topology reflecting the species relationships. This prediction was verified with grass PEPC: species relationships deduced from third positions and intron topology (Figure 2) were congruent with accepted species trees [3–5]. In this tree, the different *ppc-C4* lineages grouped into supported clusters with *ppc-B2* of related species (Figure 2). The species tree is thus recovered when nearly neutral sites are used. This pattern could not be imputed to codon usage bias between the different lineages because codon frequencies were approximately constant across the phylogenetic tree (data not shown).

The bias observed in the topology obtained with all positions suggests that a proportion of amino acids essential for the C_4 function converged between the different *ppc-C4* lineages, as confirmed by the positive selection analyses (see below). When the 21 codons under positive selection were removed from the phylogenetic analyses, a topology congruent with the species tree was obtained on the 421 remaining codons (Supplemental Data available online). This result confirmed that the clustering of *ppc-C4* was in a large part due to these few codons under positive selection.

A phylogenetic bias resulting from molecular convergence was already proposed for other genes in other organisms [6, 7]. However, these studies were not able to recover the species tree from the coding sequence. Thus, the phylogenetic bias they observed could be due

*Correspondence: pascal-antoine.christin@unil.ch (P.-A.C.), guillaume.besnard@unil.ch (G.B.)

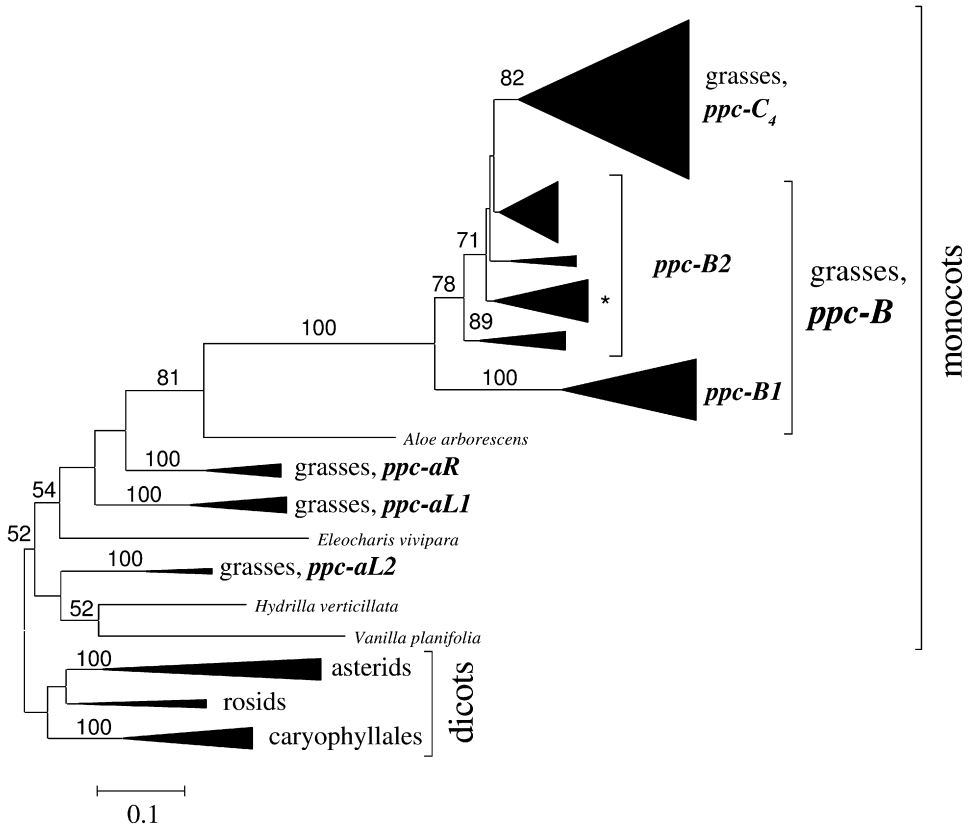

Figure 1. Maximum Likelihood Tree Containing All Grass and Main Monocot and Dicot Genes Encoding PEPC

This tree was constructed on nucleotide coding sequence with PhyML under a GTR+I+G model. Genes belonging to the different grass gene lineages and the main dicot clades are compressed. Uncompressed tree is available in Supplemental Data. Support values of 100 bootstrap replicates are indicated above branches when greater than 50%. The position of *Centropodia forskalii* gene with a serine at position 780 is indicated by an asterisk. The logarithm of the likelihood for this tree was −56730.88.

to complex gene evolutionary history (e.g., exon or gene transfer). Our study is the first to clearly show that phylogenetic reconstruction methods can be misleading because of a small proportion of convergent codons. It also highlights the importance of understanding how different parts of the data influence phylogenetic reconstruction. If using all codon positions increases the number of characters and thus the accuracy of the tree constructions, third positions and introns can, under certain conditions, better represent the true evolutionary history of the genes [8].

C_4 PEPC Evolved through Parallel Changes

It was shown before that an alanine amino acid conserved in all the known non-C_4 PEPCs changed to a serine in all C_4 PEPCs (position 780 in *Zea mays*, CAA33317) [9], representing a strong example of parallel changes. Other modifications of the coding sequence are expected because C_4 PEPCs present catalytic properties and sensitivities toward repressors different from non-C_4 isoforms [10, 11]. In order to identify sites that underwent adaptative changes during C_4 evolution in grasses, we performed positive selection tests that use a ω (*dN/dS* ratio) greater than 1 as evidence of past positive selection [12, 13]. Different codon models were optimized on third positions and intron topology and compared with likelihood ratio tests. The model allowing a proportion of codons to evolve under positive selection in branches defined a priori as the foreground branches (in this case, branches leading to *ppc-C_4*, identified by an alanine-to-serine transition at position 780) was significantly better than the null models (A versus M1a, df = 2, p value < 0.0001; A versus A′, df = 1, p value = 0.066; see Experimental Procedures for further details on the models). This result shows that *ppc-C_4* evolved under adaptive molecular evolution. To take into account the uncertainty in topology, the three codon models were run again with 11 alternative topologies sampled during the Bayesian search. 21 out of the 442 codons considered (4.8%) were identified as having evolved under positive selection in branches leading to *ppc-C_4* with a posterior probability greater than 0.95 in all analyses (Figure 3). By changing the nominal value of the test to 0.99 and 0.999, the number of codons under positive selection is reduced to 15 (3.4%) and 12 (2.7%), respectively (Figure 3).

These results show that the same positions evolved under positive selection in the different grass *ppc-C_4* lineages. Many of these sites were mutated recurrently to an identical amino acid (Figure 3). In addition, some of the amino acids under positive selection identified in grasses underwent the same transitions between non-C_4 and C_4 PEPC in other C_4 systems, even in very distant families such as Asteraceae or Amaranthaceae (Figure 3). In addition to the alanine-to-serine transition at position 780, amino acids at positions 517, 577, 665,

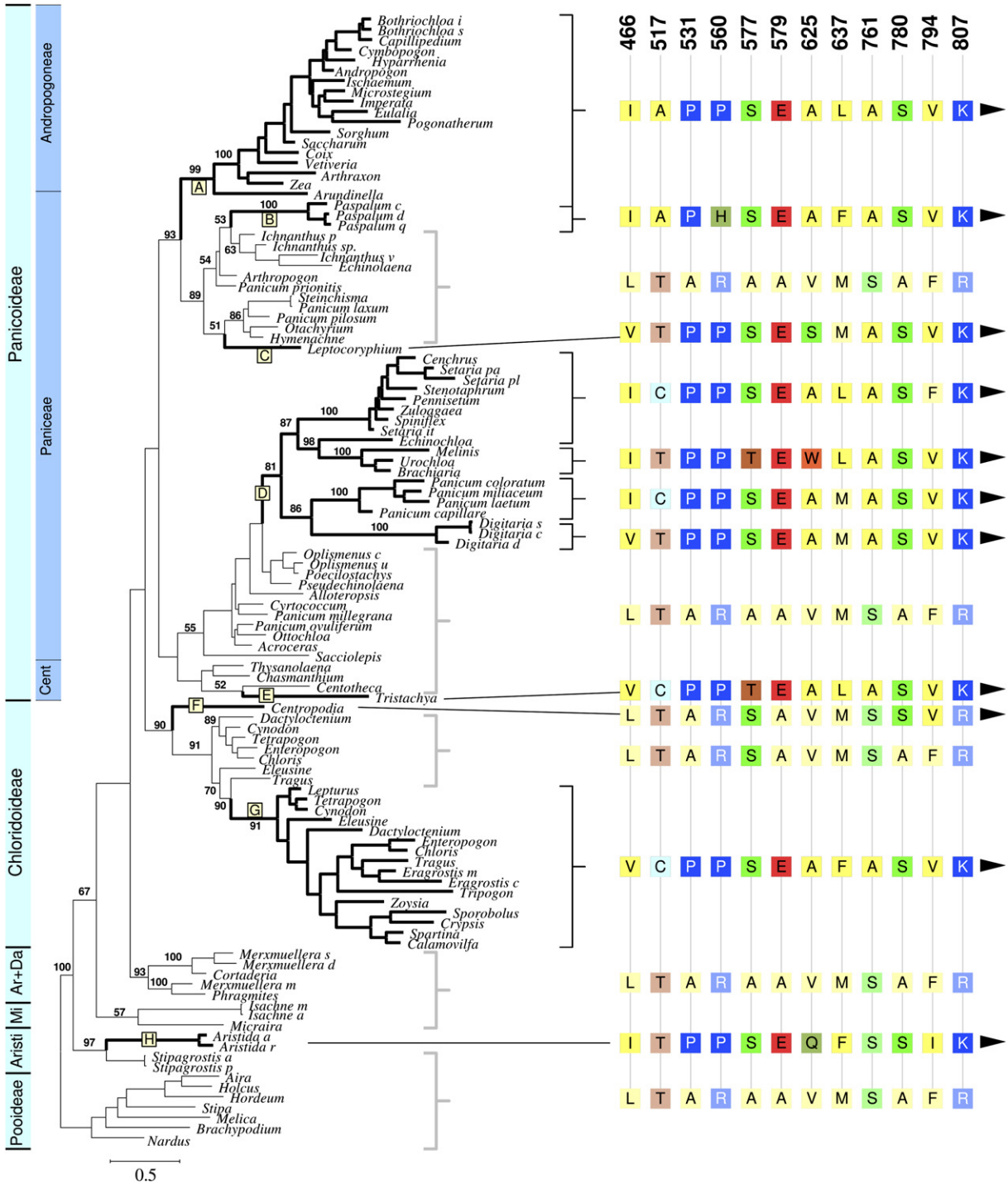

Figure 2. Bayesian Tree Constructed with MrBayes on Third Positions and Introns Combined, Including *ppc-B2* and *ppc-C_4* Genes

Branches leading to *ppc-C_4*, determined by the presence of a serine at position 780, are in bold. Capital letters identify branches used in the positive selection tests. Bayesian support values greater than 0.5 are indicated for the principal branches. Support values for all branches are available in Supplemental Data. Subfamilies are indicated on the left of the tree. The three main Panicoideae tribes, Andropogoneae, Paniceae, and Centotheceae, are indicated. Aristi, Aristidoideae; Mi, Micrairoideae; Ar+Da, Arundinoideae + Danthonioideae; Cent, Centotheceae. In some Chloridoideae, both *ppc-B2* and *ppc-C_4* are present, suggesting an ancestral gene-duplication event. On the right, the most frequent amino acid of each clade is shown for the 12 sites (positions indicated correspond to *Zea mays* PEPC; CAA33317) with a posterior probability greater than 0.999 of having evolved under positive selection in branches leading to *ppc-C_4* (Figure 3). Black triangles on the right indicate *ppc-C_4* lineages. Residues with similar biochemical properties are identically colored. For visual clarity, C_3-specific amino acids are brightened.

	Species / Clade	n	Codons under positive selection: PP > 0.999												PP > 0.99			PP > 0.95					
			466	517	531	560	577	579	625	637	761	780	794	807	572	599	813	502	596	665	733	863	866
Grasses *ppc-B2*	main	40	LVI	T	A	R	A	A	V	M	S	A	F	R	EV	IT	R	GA	E	HNRL	FVL	NED	DE
	Chloridoideae	7	L	T	A	R	SA	AE	V	M	S	A	F	R	EQI	I	RK	GP	E	N	F	DSN	ED
	Panicum prionitis	1	I	T	A	R	S	E	V	M	S	A	V	R	E	I	R	G	E	H	F	N	D
	Alloteropsis	1	L	A	A	R	A	A	V	M	S	A	F	R	Q	I	R	G	E	H	F	N	D
Grasses *ppc-C_4*	A - *Arundinella*	1	I	A	P	R	S	E	A	F	A	S	F	K	Q	V	K	A	E	N	V	N	E
	A - Andropogoneae	17	ILV	A	P	P	S	E	A	L	A	S	V	KR	Q	V	K	P	Q	N	VF	K	E
	B - *Paspalum*	3	I	A	P	H	S	E	A	F	A	S	V	K	Q	V	R	A	M	N	F	K	ED
	C - *Leptocoryphium*	1	V	T	P	P	S	E	S	M	A	S	V	K	Q	V	K	S	Q	N	F	K	D
	D - *Setaria* clade	8	I	C	P	P	S	E	AG	LM	A	S	FV	K	Q	V	K	S	Q	N	M	N	E
	D - *Echinochloa*	1	I	C	P	P	S	E	A	F	A	S	F	K	Q	V	K	T	Q	N	F	N	D
	D - *Urochloa* clade	3	I	T	P	P	T	E	WL	L	A	S	V	K	Q	VI	K	S	Q	N	F	DE	DE
	D - *Panicum* clade	4	I	C	P	P	S	E	A	M	A	S	V	K	Q	I	K	A	QR	N	VF	K	EQ
	D - *Digitaria* clade	3	V	T	P	P	S	E	A	M	A	S	V	K	Q	I	K	A	Q	N	F	K	E
	E - *Tristachya*	1	V	C	P	P	T	E	A	L	A	S	V	K	E	I	R	T	Q	N	F	N	E
	F - *Centropodia*	1	L	T	A	R	S	A	V	I	S	S	V	R	Q	V	K	G	E	R	L	K	D
	G - Chloridoideae	16	VI	CT	P	P	SA	E	AV	FL	A	S	V	KN	Q	V	K	V	Q	N	VF	K	E
	H - *Aristida*	2	I	T	P	P	S	E	Q	F	S	S	I	K	Q	V	R	A	M	H	M	K	E
Non grasses	*Eleocharis* (C_4)	1	L	A	P	R	V	T	A	L	S	S	F	R	K	I	R	P	E	N	F	N	E
	Alternanthera f (C_3)	1	L	T	A	R	A	T	I	M	S	A	F	R	E	I	K	G	E	H	F	S	E
	Alternanthera s (C_3-C_4)	1	M	T	A	R	S	T	I	M	S	A	F	R	E	I	K	G	E	H	F	S	E
	Alternanthera p (C_4)	1	M	A	A	R	S	T	I	M	A	S	F	G	E	I	K	G	E	N	F	S	E
	Amaranthus (C_4)	1	M	T	A	R	S	T	I	M	A	S	C	K	Q	I	K	G	E	H	L	S	E
	Suaeda l (C_3)	1	M	T	A	R	S	T	I	M	S	A	F	K	E	I	E	G	E	H	L	D	E
	Suaeda e (C_4)	1	M	T	A	R	S	T	I	M	A	S	F	P	K	I	E	G	E	H	F	D	E
	Suaeda a (C_4)	1	M	T	A	R	S	T	V	M	S	S	C	K	E	I	E	G	E	N	L	D	E
	Flaveria pr (C_3)	1	L	T	A	R	A	A	I	I	S	A	F	K	E	I	Q	G	E	H	F	N	E
	Flaveria br (C_4-like)	1	L	T	A	R	S	A	I	I	S	A	F	K	E	I	Q	G	E	H	F	N	E
	Flaveria tr (C_4)	1	L	T	A	R	A	T	V	I	S	S	F	K	E	I	Q	G	E	N	F	N	E

Figure 3. Amino Acids Detected as Evolving under Positive Selection in Branches Leading to Genes Encoding C_4 PEPC in Grasses

These sites were detected with a posterior probability (PP) greater than 0.999, 0.99, or 0.95. The amino acids are shown for the different C_4 and non-C_4 PEPC gene lineages (capital letters identify independent grass *ppc-C_4* lineages as identified on Figure 2). When one lineage exhibited different amino acids, the most abundant is written first. For grasses, the number of sequences included in each lineage is indicated (n), as is the photosynthetic type for nongrasses. Amino acids that differ between *ppc-B2* and *ppc-C_4* are highlighted in blue. Amino acids that underwent the same changes in non-grass C_4 PEPCs are in green.

and 761 recurrently changed from the same C_3 residue to identical C_4-specific amino acid in grasses and other families (Figure 3). The evolution of a C_4-specific PEPC was performed through many parallel changes in a high number of independent C_4 lineages, highlighting the repeatability of some evolutionary processes.

Phenotypic convergence between distant lineages is a widespread feature and concerns morphological as well as physiological traits. The recurrent appearance of the same phenotype through convergent molecular evolution has already been demonstrated [14–20]. Some studies traced the convergence to different modifications of the same gene [15–17, 19] or to the same mutations taking place independently in different lineages [14, 18, 20]. However, these cases concerned only a small proportion of sites in a restricted number of lineages. Our study reports the first case of such a high level of molecular convergent evolution in up to eight distinct lineages. The observed amino acid transitions between non-C_4 and C_4 PEPC enzymes are all due to a single nucleotide change. This increases the probability of these mutations occurring by chance. The mutations that improve the encoded enzyme can later be fixed by natural selection. The presence of a non-C_4 PEPC gene (i.e., *ppc-B2*) with a nucleotide sequence allowing the acquisition of C_4-advantageous amino acids through simple single nucleotide changes likely favored recurrent evolution of the C_4 pathway by allowing a rapid and efficient acquisition of a C_4-specific PEPC, the key enzyme of this photosynthetic pathway.

The sites under selection show different degrees of parallelisms. For instance, residues at positions 531, 579, 761, 780, and 807 mutated to an identical amino acid in six to eight grass C_4 PEPC lineages (parallel changes sensu stricto; Figures 2 and 3). In contrast, residues at positions 502, 596, and 625 changed to a different amino acid (Figures 2 and 3), suggesting that the C_4 characteristics are conferred by the absence of the non-C_4 amino acid at these positions rather than the presence of a C_4-specific amino acid. Although the latter does not match the strict definition of parallel change, it corresponds to parallel genotypic adaptation [21] because the same locus (i.e., *ppc-B2*) evolved independently through similar changes to fulfil the same function (atmospheric CO_2 fixation in mesophyll cells). Unfortunately, the effects of these different changes are difficult to predict because the described active sites and regulation targets of the PEPC [22, 23] are not affected. The alanine-to-serine transition (position 780, Figure 3) has been shown to alter the catalytic properties of the encoded enzyme [9, 11]. The histidine-to-asparagine transition that occurred at position 665 in C_4 grasses as well as in several C_4 dicots (Figure 3) could have an important effect on protein folding because it creates a putative N-glycosylation site (positions 665–668 [24]) that is absent from non-C_4 PEPCs. Serine at position 761 is part of

a predicted casein kinase II phosphorylation site (positions 761–763 [24]) that disappears once this serine is mutated to an alanine, which is the case in C_4 PEPCs. Breaking this phosphorylation site could have helped the acquisition of the C_4-specific regulation pattern of the PEPC. This amino acid is also part of a putative N-myristylation site (positions 757–762 [24]), which works with either an alanine or a serine. Thus, the only single-nucleotide change that is able to break the phosphorylation site without altering the myristylation site was precisely a serine-to-alanine substitution (serine in non-C_4 PEPC is encoded by a UCN codon). The effect of the other mutations is still unpredictable. The use of the 3D structure predictions could help evaluate whether some C_4-specific amino acids can putatively alter the enzyme structure and thus its catalytic properties [23].

Implications for Bioengineering

21 amino acids were detected, with high probability, to have undergone positive selection along the branches leading to grass *ppc-C4* and purifying or neutral selection in other branches. These changes are thus likely to be important for the C_4 function of the encoded enzyme. Their recurrent evolution in different lineages strongly supports their high adaptive significance, a fact that is reinforced by the similar or identical changes occurring at the same residues in very distant plant families (Figure 3). Knowledge of these C_4 putative determinants opens promising opportunities for the molecular engineering of grass C_3 crops, such as rice, barley, and wheat. This is especially relevant for the biotechnological efforts to incorporate some C_4 characteristics in C_3 crops [25–27]. Identification of the major C_4 determinants has been performed in *Flaveria* through expression of chimerical enzymes and analysis of their catalytic properties [9]. This approach allowed the detection of the alanine-to-serine transition (position 780 in maize). However, such a procedure can identify only changes having a detectable effect on the phenotype. The changes evidenced in our study have certainly minor independent effects, but, taken together, would help the optimization of C_4 function. Testing these many residues is not feasible in an experimental framework. The use of phylogenetic inference to detect potential residues important for the function of an enzyme is thus a feasible alternative and powerful approach that should be extended to other important enzymes.

Experimental Procedures

Amplification of PEPC Genes

Samples from 111 grass species were taken, focusing on the PACCAD clade that contains all C_4 grass species [1, 4]. Panicoideae, which contains several putatively independent C_4 lineages, was especially densely sampled. DNAs (listed in Supplemental Data) were obtained either from aliquots provided by other teams or extracted from leaves dried in silica gel via the CTAB method. The photosynthetic type was attributed to each species according to the literature.

Genes encoding PEPC were obtained from genomic DNA via polymerase chain reaction (PCR). The primers were designed to amplify a segment of *ppc-C4* genes as well as *ppc-B1* gene previously detected in *Oryza sativa* [28]. Because of the length of the complete gene (more than 6000 bp in *Zea mays*, X15239), we focused on a segment from exon 8 (PEPC-1362-For: 5′-CATCCGGCAGGAGTCG GAGCG-3′) to exon 10 (PEPC-2701-Rev: 5′-TGTASGCCTGGMAC ACGTTCAG-3′) that carries major C_4 determinants [9]. The PCR reaction mixture contained ~100 ng of genomic DNA template, 5 μl of 10X AccuPrime PCR Buffer II, 200 pmol of each dNTP, 20 pmol of each primer, 3 mmol of $MgSO_4$, 2.5 μl (5% vol) of DMSO, and 1 unit of a proof-reading *Taq* polymerase (AccuPrime *Taq* DNA Polymerase High Fidelity, Invitrogen) in a total volume of 50 μl. The samples were incubated for 2 min at 94°C, followed by 35 cycles consisting of 30 s at 94°C, 30 s at 57°C (annealing temperature), and 2 min at 68°C. The last cycle was followed by a 20 min extension at 68°C. Total PCR products were purified with QIAquick Gel Extraction Kit (QIAGEN). To separate the different genes (or alleles) putatively amplified, purified PCR products were cloned into the pTZ57R/T vector with InsT/Aclone PCR Product Cloning Kit (Fermentas) and PCR amplified with the M13 primers. Between 8 and 20 positive clones were then digested with TaqI restriction enzyme (Invitrogen). The degree of polymorphism for TaqI digestion products was high, allowing an unambiguous distinction of the different *ppc* gene lineages. For each species, inserts of each clone presenting a different restriction pattern were sequenced with the M13 primers with the Big Dye 3.1 Terminator cycle sequencing kit (Applied Biosystems), according to the provider instructions, and separated on an ABI Prism 3100 genetic analyzer (Applied Biosystems). A segment of about 1500 bp, including ~40% of the total coding sequence and two introns, was sequenced. All sequences have been deposited in the EMBL database (accession numbers in Supplemental Data).

DNA Sequence Analyses

For PEPC-gene segments isolated from genomic DNA, exons were identified by homology with *Zea mays*, *Sorghum bicolor*, and *Oryza sativa* genes (X15239, X63756, and AK101274, respectively) and according to, when possible, the GT-AG rule. Coding sequences were then translated into amino acids and aligned with ClustalW [29]. Once retranslated into nucleotides, alignment was checked visually. 19 grass and 23 nongrass *ppc* genes available on GenBank were added to the data set (Supplemental Data). A phylogenetic tree was inferred both by maximum likelihood via PhyML [30], DNAML [31], and PAUP* [32] (NNI branch swapping on 151 trees found during a first round of tree selection with 1000 random addition sequences with TBR branch swapping under the Parsimony criterion; this was needed to reduce computational time) and by Bayesian inference via MrBayes [33] (two runs of 10,000,000 generations with four chains, burn-in period of 2,000,000) under a GTR model with base frequencies gamma shape parameter and proportion of invariants estimated from the data (hereafter referred as GTR+I+G). PhyML [30] and ProML [31] were further used to compute a phylogenetic tree based on the amino acids sequence under a JTT substitution model with a gamma shape parameter. For DNAML and ProML, gamma shape parameter was fixed to the value estimated by PhyML. These analyses allowed the identification of the number of gene lineages present in grasses and their relationships to each other.

Further analyses included only *ppc-C4* lineage and its closest non-C_4 ancestor (hereafter named *ppc-B2*, Figure 1). More distantly related sequences were omitted to avoid saturation of fast-evolving nucleotides such as introns and third positions. To distinguish the phylogenetic information provided by the different parts of the sequences, two data sets were created. First, all coding positions were considered for a total of 1326 bp. Second, the third codon positions (442 bp) were combined with introns 8 and 9. The introns were extracted and aligned with ClustalW with gap opening and gap extension penalties set to 15 and 6.6 for the pairwise and multiple alignments. To avoid subjectivity, intron alignments were checked visually but not manually edited. Because of their fast evolutionary rate, introns are useless to resolve basal nodes but give a strong signal to infer the top nodes. Their use in combination with unequivocally aligned third positions appeared as the best way to obtain a supported tree only weakly affected by selective pressures. The substitution model used for the introns was the HKY model. All coding positions and third positions of codons were analyzed under a GTR+I+G model. Best-fit substitution models were determined with hierarchical likelihood ratio tests (LRT). Both data sets were analyzed by Bayesian inference with MrBayes 3.1 [33]. Each analysis was run twice for 10,000,000 generations. Sample frequency was set to 1000 generations. Prior distributions were left to their default

values. The number of chains in each run was increased from four for all coding positions analysis to six for introns and third position analysis because of convergence problems. Base frequency, which was the only parameter common to the substitution models of these two data sets, was optimized separately for each partition (option unlink in MrBayes).

Selection Tests

To test for the action of positive selection at particular sites of the nucleotide sequence of the *ppc* genes along branches leading to *ppc-C4*, three different codon models [12, 13] were optimized on the topology obtained by combining third codon positions and introns via codeml [34]. The neutral model M1a allows ω (the *dN/dS* ratio) to vary among codons. This parameter is constant among the branches of the tree and its value is allowed to be either 1 (neutral) or smaller than 1 (purifying selection). The alternative model, model A, allows ω to vary among both sites and branches. It requires the specification of two branch types, the background branches in which selective pressures are either neutral or purifying and the foreground branches under positive selection ($\omega > 1$). The last model A′ allows ω to vary among sites and branches but not to be greater than 1. It is therefore identical to model A except that the ω value in foreground branches is fixed to 1 for sites that differ between foreground and background branches. Models were compared with LRT. Test 1 compares model M1a and A and thus tests for the occurrence of different selective pressures on the foreground branches [13]. Test 2, which compares models A′ and A, is more conservative and specifically tests the significance of a ω value greater than 1 on the foreground branches [13].

Models A′ and A require an a priori identification of the foreground branches. All the branches leading to full C_4 PEPC groups (identified by a serine at position 780, see Figure 2) were used simultaneously as foreground branches. To ensure that the results were not due to a bias in the tree used, the same procedure was repeated with topologies sampled during the Bayesian search. Trees were taken each 500,000 generations between 5,000,000 and 10,000,000 for a total of 11 additional topologies. By the Bayes Empirical Bayes approach [35], only codons with posterior probability of being under positive selection greater than a given threshold (i.e., 0.95, 0.99, or 0.999) in all 12 analyses were considered as having evolved under positive selection during C_4 evolution.

The most likely ancestral residue at position 780 was determined with codeml under a F3x4 model of codon substitution. The deduced amino acid was used to trace the *ppc-C4* evolution events on the phylogenetic trees.

Subsequent to these analyses, sequences from nongrasses *ppc-C4* and their related non-C_4 PEPC gene available in GenBank were aligned to the grass DNA sequences. The amino acids corresponding to sites under positive selection in grasses were reported.

Supplemental Data

Five figures and two tables are available at http://www.current-biology.com/cgi/content/full/17/14/1241/DC1/.

Acknowledgments

This work was funded by Swiss NSF grant 3100AO-105886/1. N.S. and V.S. were funded by the European Commission (Marie Curie EST "HOTSPOTS," contract MEST-CT-2005-020561). We thank the Swiss Institute of Bioinformatics for access to the Vital-IT cluster. The authors are especially thankful to F. Anthelme, Y. Bouchenak-Khelladi, V.R. Clark, M. Gonzalez, T.R. Hodkinson, J. Kissling, C. Lavergne, A. Persico, T. Renaud, P. Rondeau, S. Sunkkaew, A. Teerawatanakon, and Y. Wang who provided either DNA aliquots or grass samples. N. Fumeaux at the herbarium of the botanical garden of Geneva helped with grass identification. Finally, O. Brönnimann, L. Büchi, M. Chapuisat, P.B. Pearman, E. Samaritani, and I. Sanders made useful comments on the earlier versions of the manuscript. We thank two anonymous reviewers for useful comments.

Received: May 3, 2007
Revised: June 4, 2007
Accepted: June 12, 2007
Published online: July 5, 2007

References

1. Sage, R.F. (2004). The evolution of C_4 photosynthesis. New Phytol. *161*, 341–370.
2. Lepiniec, L., Vidal, J., Chollet, R., Gadal, P., and Crétin, C. (1994). Phospho*enol*pyruvate carboxylase—structure, regulation and evolution. Plant Sci. *99*, 111–124.
3. Giussani, L., Cota-Sánchez, J.H., Zuloaga, F., and Kellogg, E.A. (2001). A molecular phylogeny of the grass subfamily Panicoideae (Poaceae) shows multiple origins of C_4 photosynthesis. Am. J. Bot. *88*, 1993–2012.
4. GPWG-Grass Phylogeny Working Group (2001). Phylogeny and subfamilial classification of the grasses (Poaceae). Ann. Mo. Bot. Gard. *88*, 373–457.
5. Sanchez-Ken, J.G., Clark, L.G., Kellogg, E.A., and Kay, E.E. (2007). Reinstatement and emendation of subfamily Micrairoideae (Poaceae). Syst. Bot. *32*, 71–80.
6. Stewart, C.B., Schilling, J.W., and Wilson, A.C. (1987). Adaptive evolution in the stomach lysozymes of foregut fermenters. Nature *330*, 401–404.
7. Kriener, K., O'hUigin, C., Tichy, H., and Klein, J. (2000). Convergent evolution of major histocompatibility complex molecules in humans and New World monkeys. Immunogenetics *51*, 169–178.
8. Savolainen, V., Chase, M.W., Salamin, N., Soltis, D.E., Soltis, P.E., Lopez, A.J., Fedrigo, O., and Naylor, G.J.P. (2002). Phylogeny reconstruction and functional constraints in organellar genomes: plastid *atpB* and *rbcL* sequences versus animal mitochondrion. Syst. Biol. *51*, 638–647.
9. Bläsing, O.E., Westhoff, P., and Svensson, P. (2000). Evolution of C_4 phospho*enol*pyruvate carboxylase in *Flaveria*, a conserved serine residue in the carboxyl-terminal part of the enzyme is a major determinant for C_4-specific characteristics. J. Biol. Chem. *275*, 27917–27923.
10. Dong, L.Y., Masuda, T., Kawamura, T., Hata, S., and Izui, K. (1998). Cloning, expression, and characterization of a root-form phospho*enol*pyruvate carboxylase from *Zea mays*: comparison with the C4-form enzyme. Plant Cell Physiol. *39*, 865–873.
11. Svensson, P., Bläsing, O.E., and Westhoff, P. (2003). Evolution of C_4 phospho*enol*pyruvate carboxylase. Arch. Biochem. Biophys. *414*, 180–188.
12. Yang, Z.H., and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. *19*, 908–917.
13. Zhang, J.Z., Nielsen, R., and Yang, Z.H. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. *22*, 2472–2479.
14. Andreev, D., Kreitman, M., Phillips, T.W., Beeman, R.W., and Ffrench-Constant, R.H. (1999). Multiple origins of cyclodiene insecticide resistance in *Tribolium castaneum* (Coleoptera: Tenebrionidae). J. Mol. Evol. *48*, 615–624.
15. Mundy, N.I., Badcock, N.S., Hart, T., Scribner, K., Janssen, K., and Nadeau, N.J. (2004). Conserved genetic basis of a quantitative plumage trait involved in mate choice. Science *303*, 1870–1873.
16. Mundy, N.I. (2005). A window on the genetics of evolution: MC1R and plumage colouration in birds. Proc. R. Soc. Lond. B. Biol. Sci. *272*, 1633–1640.
17. Protas, M.E., Hersey, C., Kochanek, D., Zhou, Y., Wilkens, H., Jeffery, W.R., Zon, L.I., Borowsky, R., and Tabin, C.J. (2006). Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. Nat. Genet. *38*, 107–111.
18. Yokoyama, R., and Yokoyama, S. (1990). Convergent evolution of the red- and green-like visual pigment genes in fish, *Astyanax fasciatus*, and human. Proc. Natl. Acad. Sci. USA *87*, 9315–9318.
19. Zakon, H.H., Lu, Y., Zwickl, D.J., and Hillis, D.M. (2006). Sodium channel genes and the evolution of diversity in communication signals of electric fishes: convergent molecular evolution. Proc. Natl. Acad. Sci. USA *103*, 3675–3680.
20. Zhang, J.Z. (2006). Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. Nat. Genet. *38*, 819–823.
21. Wood, T.E., Burke, J.M., and Rieseberg, L.H. (2005). Parallel genotypic adaptation: when evolution repeats itself. Genetica *123*, 157–170.

22. Kai, Y., Matsumura, H., Inoue, T., Terada, K., Nagara, Y., Yoshinaga, T., Kihara, A., Tsumura, K., and Izui, K. (1999). Three-dimensional structure of phospho*enol*pyruvate carboxylase: a proposed mechanism for allosteric inhibition. Proc. Natl. Acad. Sci. USA *96*, 823–828.
23. Kai, Y., Matsumura, H., and Izui, K. (2003). Phospho*enol*pyruvate carboxylase: three-dimensional structure and molecular mechanisms. Arch. Biochem. Biophys. *414*, 170–179.
24. Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro, E., Langendijk-Genevaux, P.S., Pagni, M., and Sigrist, C.J.A. (2006). The PROSITE database. Nucleic Acids Res. *34*, D227–D230.
25. Matsuoka, M., Furbank, R.T., Fukayama, H., and Miyao, M. (2001). Molecular engineering of C_4 photosynthesis. Annu. Rev. Plant Biol. *52*, 297–314.
26. Miyao, M. (2003). Molecular evolution and genetic engineering of C_4 photosynthetic enzymes. J. Exp. Bot. *54*, 179–189.
27. Raines, C.A. (2006). Transgenic approaches to manipulate the environmental responses of the C_3 carbon fixation cycle. Plant Cell Environ. *29*, 331–339.
28. Christin, P.A., Salamin, N., Savolainen, V., and Besnard, G. (2007). A phylogenetic study of the phospho*enol*pyruvate carboxylase multigene family in *Poaceae*: understanding the molecular changes linked to C_4 photosynthesis evolution. Kew Bull. *62*, in press.
29. Thompson, J.D., Higgins, D.J., and Gibson, T.J. (1994). ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and matrix choice. Nucleic Acids Res. *22*, 4673–4680.
30. Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. *52*, 696–704.
31. Felsenstein, J. (2005). PHYLIP (Phylogeny Inference Package) version 3.6 (Seattle, WA: Department of Genome Sciences, University of Washington).
32. Swofford, D.L. (2002). PAUP*: Phylogenetic Analysis Using Parsimony (* and other methods), version 4.0b8 (Sunderland, MA: Sinauer Associates).
33. Ronquist, F., and Huelsenbeck, J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics *19*, 1572–1574.
34. Yang, Z.H. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. *13*, 555–556.
35. Yang, Z.H., Wong, W.S.W., and Nielsen, R. (2005). Bayes empirical Bayes inference of amino acids sites under positive selection. Mol. Biol. Evol. *22*, 1107–1118.

Accession Numbers

The accession numbers assigned to the sequences we submitted to GenBank are from AM689877 to AM689901 and from AM690209 to AM690312.