WILEY-BLACKWELL

Cladistics 24 (2008) 1–12

Cladistics

10.1111/j.1096-0031.2008.00236.x

Evolution of small inversions in chloroplast genome: a case study from a recurrent inversion in angiosperms

Santiago Andrés Catalano*, Beatriz Ofelia Saidman and Juan César Vilardi

Departamento de Ecología Genética y Evolución, Facultad de Ciencias Exactas y Naturales, Universidad Nacional de Buenos Aires. Intendente Güiraldes 2160, Ciudad Universitaria, C1428EGA – Capital Federal, Argentina

Accepted 7 July 2008

Abstract

Small inversions (SIs) in the chloroplast genome of angiosperms are ubiquitous. These inversions are always flanked by inverted repeats (palindromes or quasipalindromes) between approximately 8 and 50 bp long that form a hairpin structure when the DNA is single-stranded. We evaluated different methodological and empirical issues about SI evolution. As a case study, we analysed an SI recently discovered in the *psb*C–*trn*S intergenic region of *Prosopis* (Fabaceae). First, we analysed how inversions can be optimized in cases where the inverted segment also shows indels and substitutions, proposing a method based on Fixed States Optimization. Second, we evaluated the occurrence of this inversion on a phylogeny that includes the major lineages of angiosperms. Finally, we assessed whether the occurrence of this inversion was related to the thermodynamic stability of the hairpin structure (measured by its corresponding free energy) and/or the length of the palindromes by using a modified version of Maddison's Concentrated Changes Test. Hairpin structure was conserved in most of the 154 sequences analysed, with the inversion taking place at least 10 times in different lineages (monocots, magnoliids, rosids). As was previously proposed for other SIs, our analysis strongly suggests that the occurrence of this inversion is correlated with higher hairpin stability. In contrast, we found no evidence of a correlation with longer palindromes. Our results are in agreement with the hypothesis that hairpin formation is a requisite for SI occurrence. However, alternative explanations cannot be discarded.

The existence of large inversions in the chloroplast genome has been documented for more than 20 years (Palmer, 1985; Jansen and Palmer, 1987; Milligan et al., 1989; Raubeson and Jansen, 1992). More recently, a large number of small- to medium-size inversions (i.e. those that present less than a few hundred base pairs; called thereafter SIs) were reported in a diverse array of angiosperm lineages (Kelchner and Wendel, 1996; Graham and Olmstead, 2000; Mes et al., 2000; Kim and Lee, 2005; Bain and Jansen, 2006; Swangpol et al., 2007). A distinctive feature of these inversions is that they are always flanked by inverted repeats (IRs) that range from 8 to 50 bp (Fig. 1a). This region forms a stable single-stranded hairpin structure (Fig. 1b) on which the IRs form the stem and the segment between them (spacer) forms the loop. Hairpins in chloroplast genome are very common in the 3′ end of *tRNA* genes and between genes with a tail to tail configuration (Kim and Lee, 2005). These structures are thought to be related to the stabilization of the corresponding RNA molecules (Stern and Gruissem, 1987; Stern et al., 1989). Although most of the inversions associated with flanking palindromes were found in chloroplasts, they are not exclusive to this genome. Small inversions were also found in plant (Dumolin-Lapègue et al., 1998) and human mitochondrial genomes (Musumeci et al., 2000; Blakeley et al., 2006).

These SIs are generally recognized by pairwise comparisons between sequences of closely related taxa. In those analyses no algorithms are used to evaluate the occurrence of SI, although a parsimony criterion is often implicitly or explicitly invoked. For example, Kelchner and Wendel (1996, p. 260), argued that in *Chusquea*:

*Corresponding author:
E-mail address: sacatalano@gmail.com

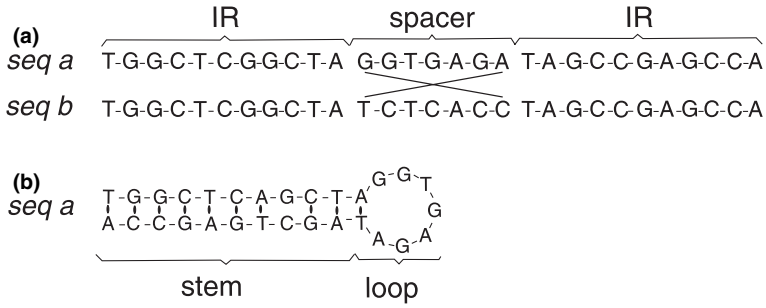

Fig. 1. (a) Example of two sequences differing by an inversion. The spacer in sequence a is the reverse complement of the spacer in sequence b. (b) Hairpin formed when sequence a is in single-stranded condition. IR, inverted repeat.

> …four adjacent nucleotide polymorphisms are more parsimoniously explained by a single mutational event, in this case small inversions…

Pairwise comparisons are suitable to analyse the existence of inversions when the number of sequences and divergence between them is low. However, when the inverted region also presents substitutions and/or indels, this task requires an algorithmic approach. Algorithms developed to evaluate inversions within a phylogenetic framework are devoted to analysing inversions at the chromosome level. These algorithms evaluate the order and sense of genes and intend to minimize the number of inversions and transpositions events that occur between two chromosomes (Moret et al., 2002). However, no algorithm has been proposed to optimize small DNA segments when their sense is not pre-established and there are substitutions and/or indels.

Kelchner and Wendel (1996) suggested that the hairpin thermodynamic stability (measure by its corresponding free energy values; ΔG) affects the occurrence of SI. According to these authors, the structures with higher hairpin stability will more likely form hairpins and, concomitantly, will have higher frequency of inversions. Kim and Lee (2005) proposed that SI occurrence may be correlated with the length of the palindromes. Nevertheless, no analyses have been performed to test these hypotheses. Their evaluation within a phylogenetic framework requires the use of a comparative method as well as a method to optimize the inversion on the tree.

In the present paper we evaluated different methodological and empirical issues about SI evolution. As a case study, we analysed an inversion recently documented in *psb*C–*trn*S chloroplast intergenic region of *Prosopis* L. (Fabaceae). In particular we evaluated: (i) how inversions can be optimized in cases where the inverted segment also contain indels and substitutions; (ii) the occurrence of this inversion throughout angiosperms; (iii) the possible association of inversion events with higher thermodynamic stability of the hairpin structure and/or longer IRs; and (iv) the effect of different coding schemes and character delimitation on the evaluation of these associations. Our analysis showed that the inversion took place at least 10 times in different lineages of angiosperms, and that the occurrence of this inversion was correlated with higher hairpin stability but not with longer IRs. The different methods used to optimize ΔG values strongly affected the results of the correlation analyses. Conversely, the results were not affected by the different methods used to optimize inversions. Our results are in agreement with those of Kelchner and Wendel (1996) suggesting that hairpin structure affects the occurrence of SI. However, alternative explanations are also possible.

Methodology

Optimization of inversions

The analysis of SIs within a phylogenetic framework has been generally restricted to its coding as a binary character (binary coding approach), where each of the two senses of the inverted sequence is represented by a state (Graham and Olmstead, 2000). We present here a method to study inversions that can also handle the occurrence of substitutions and indels in the region analysed. This approach is based on Fixed State Optimization (Wheeler, 1999). In the original version of this method, homologous contiguous stretches of nucleotides are treated as one character, and the editing cost is calculated for all possible pairs of observed "states". This new character is subsequently analysed by Sankoff optimization (Sankoff and Rousseau, 1975). To cope with inversions, in the method proposed here (Fixed State Optimization with inversions, FSI), this rearrangement is treated as another molecular event, just like Fixed State Optimization treats *indels* and substitutions. The first step of FSI involves the calculation of the editing cost between each of the observed spacer sequences using the *wave-front* step of Needleman and Wunsch (1970) algorithm. Then, a second editing cost is calculated in the same manner but comparing one of the observed spacer sequences with the reverse complement of the second spacer; the cost of the

inversion is calculated as the editing cost plus an inversion penalty. The lowest of these two values is used to build a cost matrix. Then, this compound character is optimized on the tree considering the costs previously calculated. Finally, the nodes are visited to evaluate which of the transformations implied inversions. Implementation and other details about this analysis are indicated in Appendix 1.

Phylogenetic comparison

The evaluation of whether the occurrence of inversions is related to hairpins with longer stems and/or higher thermodynamic stability (evaluated by its ΔG) is akin to the analyses that intend to establish whether changes in one character (dependent variable) are concentrated on branches with a particular state in a second character (independent variable). One of the methods most commonly use to analyse such relationships is the Concentrated Changes Test (CCT; Maddison, 1990), However this test is designed to analyse two binary characters and there is no analogous test designed to study cases where the independent variable is continuous and the response variable is binary. Consequently, we transformed ΔG and IRs length into binary variables by sorting out the values in two states and performed the analysis using a modified version of CCT. More information on this analysis can be found in Appendix 2.

Optimization of free energy and stem length

The ΔG values were optimized on the angiosperm tree in three different ways. First, ΔG values were classified in two classes: high and low considering a threshold value of –10.0 kcal/mol (see the Supporting Information, Fig. S1). Subsequently, ΔG was analysed as a binary character (Fig. 2, right). In a second strategy, ΔG was directly optimized on the tree as a continuous character (Fig. 2, left) using Farris (1970) optimization as implemented in TNT (Goloboff et al., 2003, 2006). A similar strategy was followed for IRs length. Here, two limits between *short* and *long* IRs classes were evaluated: >10 and >11 bp (Fig. S2; supplementary data). A third strategy to assign ΔG to internal nodes involved the evaluation of the ancestral secondary structures (ancestral secondary structure analysis; Fig. 3). In this case, ancestral sequences were inferred for each node and their corresponding ΔG values were calculated in the same way as for the terminals. The strategy involved the separation of the sequences in three segments: (i) the spacer, (ii) 25 bases downstream the spacer (iii) 25 bases upstream the spacer. The flanking regions were optimized on the tree with POY IV Beta v. 4.0.2398 (Varón et al., 2007) using direct optimization (DO; Wheeler, 1996) and a combination of DO and iterative

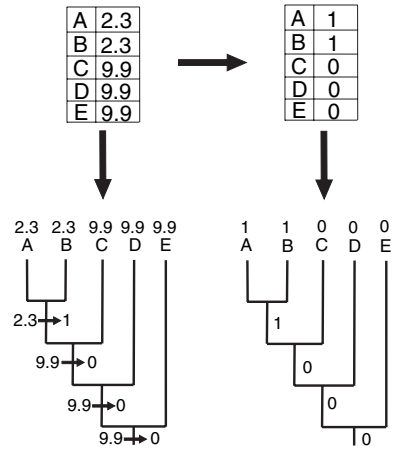

Fig. 2. Continuous (left) and binary (right) coding of ΔG values. For explanation, see text.

pass optimization (IP; Wheeler, 2003). The ancestral sequences for each node were retrieved using the command *report* (*diagnosis*). Afterwards, and for each node, flanking regions were combined with the optimal assignment of the spacer motif obtained in the FSI optimization performed in TNT. Subsequently, the secondary structure and the associated ΔG value were calculated for each ancestral sequence. These values were then forced on each node to obtain a reconstruction of the ancestral ΔG values for the whole tree[1]. The same strategy was followed to estimate ancestral values for the IR lengths.

Two different sets of costs were considered in the optimization step: Indel penalty (g) = 12, Substitution penalty (s) = 1, and $g = 2$, $s = 1$. As some sequences presented less than 25 bp downstream and upstream the spacer, a second analysis was performed considering only 17 bp, the length of the longest flanking sequences that present no missing data. Secondary structure and associated ΔG values were calculated with Quickfold (Markham and Zuker, 2005) using default options. Given the heuristics used by POY IV, and that only one optimal sequence was retrieved for each ancestral node, the optimization of the original sequences and the optimization of the reverse complement (and, afterward, its restitution to the original sequence) may potentially give different ancestral sequence assignments. This phenomenon was observed in small artificial datasets (online supplementary data). Hence, in *psb*C-*trn*S dataset we optimized both the original sequences and the reverse complements.

[1]Since these assignments were not obtained by optimizing ΔG values from the terminals, these reconstructions need not be MPRs for ΔG. See a more comprehensive development of this point in the Discussion section.

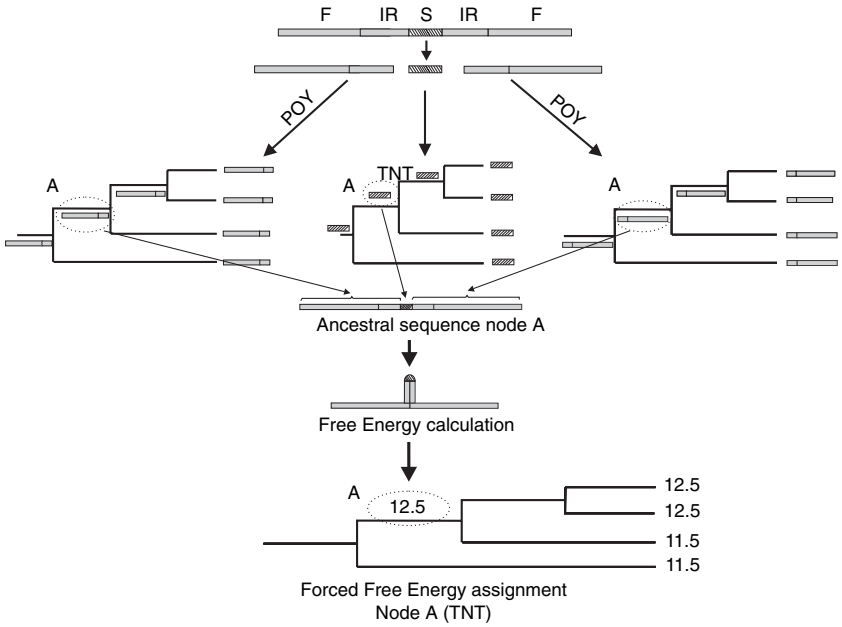

Fig. 3. Ancestral secondary structure analysis. F, flanking region; S, Spacer; IR, inverted repeat. For explanation, see text.

Dataset

The inversion studied here was originally documented in the frame of a phylogenetic analysis in the mimosoid genus *Prosopis* (Catalano et al., 2008). It is located in the large single copy (LSC) region of the plastome between *trn*S (UAG) and *psb*C genes. We downloaded all the sequences available in GenBank for this genomic region. In those species where more than one individual was retrieved, only one of them was randomly chosen to be included in the analysis. The total number of sequences analysed was 154 representing 104 angiosperm genera (Table S1). Sequences were handled with BioEdit (Hall, 1999). Two different datasets were analysed: one included all the species of each genus while the other was formed by at most two species per genus. As these analyses gave similar results, only those obtained in the two-species analysis are shown. The results of the alternative sampling are included in the Supporting Information. The backbone of the phylogenetic tree used in the phylogenetic comparison was defined by the relationships established in APG II (2003). Other studies (Olmstead et al., 1999; Barker et al., 2001; Bremer et al., 2002; Lavin et al., 2005) were also employed to define the relationships within particular clades. The resulting topology is included in the supplementary data (Fig. S3). In addition, an alternative topology with monocots as sister clade of eudicots was also considered in the analysis. The results obtained with both topologies were very similar (see the Supporting Information).

Results

Hairpin characterization

The hairpin structure present in *psb*C–*trn*S intergenic region was conserved in most of the species analysed. However, some species lacked the stem–loop structure or presented a structure that was strongly modified by several substitutions and rearrangements (Table S1). Among the sequences considered for the analyses, the length of the spacer varied from 8 to 11 bp while the IR lengths ranged from 9 to 17 bp. Nonetheless, 97% of the species presented IRs between 10 and 12 bp, and 63% presented stems of 11 bp. Most of the sequences presented ΔG values within an interval of –8.8 and –13.3 kcal/mol (Table S1).

Optimization of inversions, ΔG *and IR lengths*

Optimization analyses indicated that the inversion in *psb*C–*trn*S intergenic region occurred at least 10 times in angiosperm history (Fig. 4, Table 1). Both binary coding and FSI analysis inferred inversions on the same branches at the tips of the trees, but presented some differences when the inversions occurred on deep branches.

The reconstructions of ancestral values of ΔG using binary and continuous coding were very similar. However, the results obtained in the analysis of ancestral secondary structure differed substantially from the former analyses. The cause of this

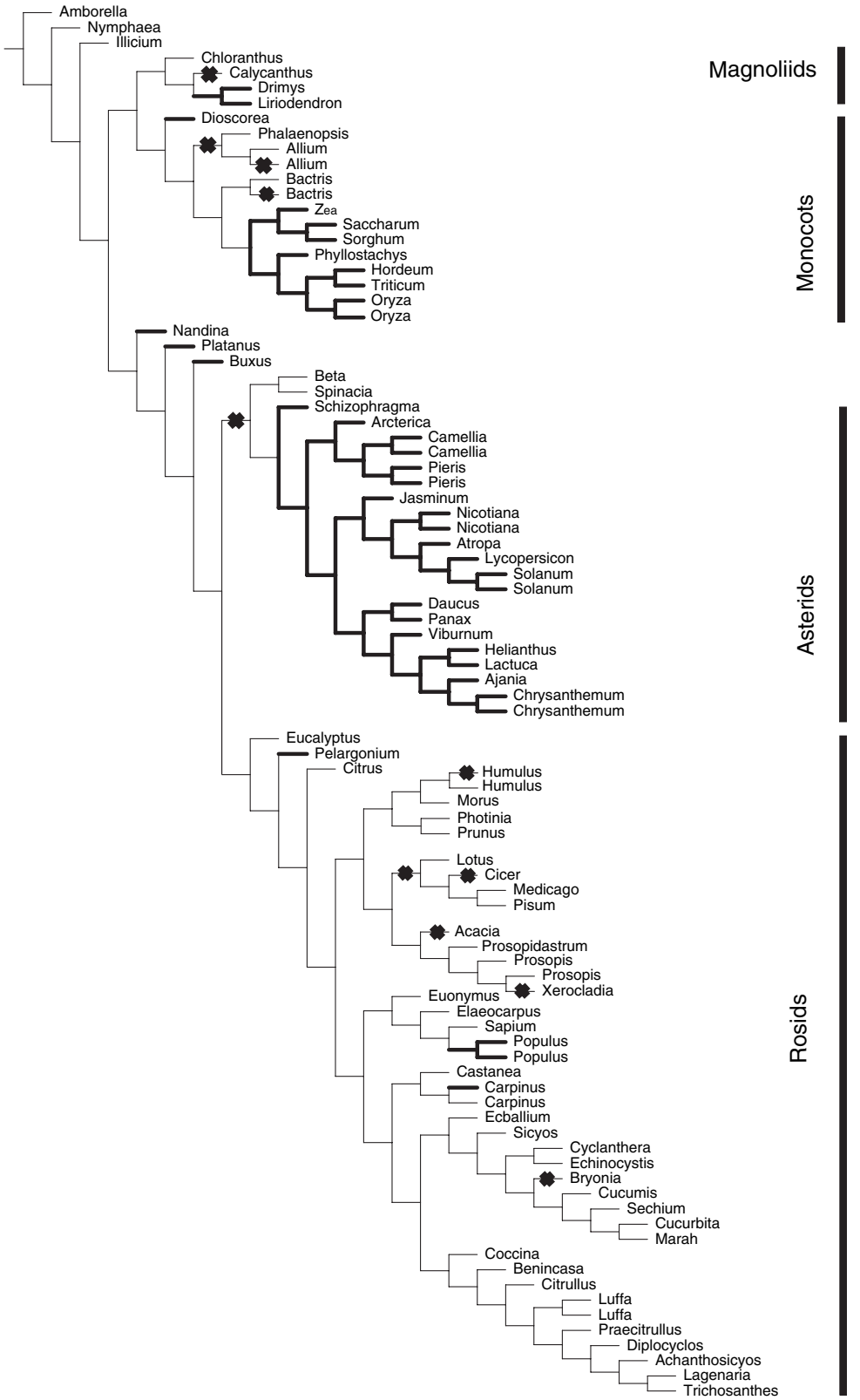

Fig. 4. One of the most parsimonious reconstructions of psbC-trnS inversion (crosses) inferred with FSI method. Thin and thick branches indicate low and high values of ΔG respectively. The topology is one of those considered in the comparative method study. Supra-ordinal names and limits follow APG II (2002).

Table 1
Relationship between hairpin stability (ΔG) and inversion occurrence. The range of values for each cell represents minimum and maximum values obtained under different reconstructions and topologies

Optimization		Number of branches		Number of inversions		Probability	
ΔG	Inversion	High ΔG	Low ΔG	In high ΔG branches	In total branches	Hyper	Simulation
Binary	Binary	102–112	58–68	0–2	11–12	0.005–0.091	0.002–0.066
	FSI	102–112	58–68	0–3	10–13	0.003–0.169	0.001–0.125
Continuous	Binary	102–112	58–68	0–2	11–12	0.005–0.091	0.002–0.130
	FSI	102–112	58–68	0–3	11–13	0.003–0.169	0.001–0.127
Sequences*	FSI	113	56	0	11	0.004	0.004

*Ancestral secondary structure analysis. Values calculated for the combination of topology-inversion reconstruction that most contradicts the hypothesis evaluated. Hyper = probabilities calculated using a cumulative hypergeometric distribution

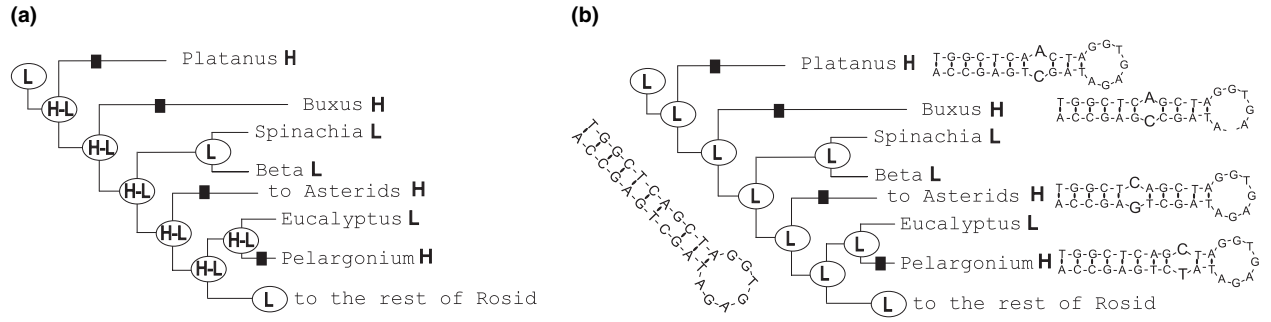

Fig. 5. Assignments of ancestral ΔG values to some deeper nodes. (a) Ancestral values obtained by optimizing ΔG values from the terminals. The presence of several terminals with high ΔG values leads to include this state in the most parsimonious state set of the intermediate nodes (b) ΔG values derived from ancestral sequences (ancestral secondary structure analysis; see Fig. 4). This reconstruction shows that ΔG high values were independently derived by different mutations (black bars) in the inverted repeats, and that the ancestral condition was low ΔG (i.e. high hairpin stability) even if many descendants have high ΔG values. The topology is one of those considered in the comparative method.

discrepancy is illustrated in Fig. 5. In binary and continuous optimizations, some of the deeper nodes presented ambiguous assignments (Fig. 5a), and one of the most parsimonious reconstructions (MPR) implied high ΔG values for all these nodes. Conversely, when the sequences themselves were optimized and ΔG values were calculated from the ancestral sequences, these ancestral nodes presented low ΔG values (i.e. high thermodynamic stability) with each lineage independently deriving high ΔG values by mutations in different parts of the stem (Fig. 5b). A similar situation occurred to that of the MRCA of *Drimys* and *Liriodendron*, where high ΔG values were independently derived in both lineages (Fig. 4). Contrasting with ΔG analyses, the different methods employed to assigned IRs length to inner nodes (i.e. binary, continuous, and ancestral sequence analysis) gave the same result.

The results of the ancestral sequence optimization were affected by: (i) different cost regimes; (ii) different optimization algorithms (DO alone or DO + IP); (iii) length of the segment analysed; and (iv) sense of the sequence optimized (the original sequences or the corresponding reverse complements). However, these parameters only changed the ancestral sequences in a region of the intergene that do not take part of the hairpin structure (i.e. downstream or upstream of the IRs). Consequently, the changes in these conditions did not alter the values of free energy or stem length calculated for ancestral structures.

Relationship between inversion occurrence and IR free energy/length

Free energy. The ambiguities in binary and continuous ΔG optimization strongly affected the significance of the correlation between ΔG and the occurrence of inversions. The probabilities differed up to 55 times among reconstructions (Table 1). The main cause for these differences was that the incorrect ancestral ΔG estimation previously outlined (Fig. 5) produced high P values (non-significant). When the ancestral ΔG values were obtained by evaluating the ancestral secondary structure, the phylogenetic comparison strongly suggested that the occurrence of inversions was correlated with higher hairpin stability (Table 1), with all the inversions situated on branches of high hairpin stability. The results scarcely differed among the different topologies evaluated.

The different cost regimes and methods to optimize inversions (binary coding or FSI) did not affect the

Table 2
Relationship between hairpin stability and inversion occurrence considering only species with IRs of 11 bp. The range of values for each cell represents minimum and maximum values obtained under different reconstructions and topologies

Optimization		Number of branches		Number of inversions		Probability	
Energy	Inversion	High ΔG	Low ΔG	In high ΔG branches	In total branches	Hyper	Simulation
Binary	Binary	48–54	78–84	0–1	7–8	0.029–0.140	0.021–0.118
	FSI	48–54	78–84	0–1	7–8	0.024–0.140	0.012–0.110
Continuous	Binary	48–54	78–84	0–1	7–8	0.023–0.140	0.015–0.114
	FSI	48–54	78–84	0–1	7–8	0.023–0.140	0.016–0.113
Sequences*	FSI	48	84	0	8	0.023	0.033

*Ancestral secondary structure analysis. Values calculated for the combination of topology-inversion reconstruction that most contradicts the hypothesis evaluated. Hyper = probabilities calculated using a cumulative hypergeometric distribution

Table 3
Relationship between IRs length and inversion occurrence. The range of values for each cell represents minimum and maximum values obtained under different reconstructions and topologies. *Limit* represents the threshold between short and long IRs

Optimization			Number of branches		Number of inversions		Probability	
Length	Limit	Inversion	Short IRs	Long IRs	In short IRs	In total branches	Hyper	Simulation
Binary	>10	Binary	11–12	162–166	2	8–12	0.960–0.990	0.974–0.999
		FSI	11–13	162–166	2	8–12	0.951–0.991	0.968–0.998
	>11	Binary	158–164	9–11	9–11	10–12	0.123–0.698	0.123–0.752
		FSI	158–164	10–12	9–11	11–13	0.144–0.728	0.150–0.791
Continuous	>10	Binary	11–12	162–166	1–2	8–12	0.950–0.990	0.978–0.999
		FSI	11–13	162–166	1–2	8–12	0.960–0.990	0.972–0.998
	>11	Binary	158–164	9–11	9–11	10–12	0.120–0.690	0.123–0.756
		FSI	158–164	10–12	9–11	11–13	0.144–0.690	0.147–0.759

Hyper = probabilities calculates using a cumulative hypergeometric distribution.

results of the comparative analysis. The probabilities calculated when the null distribution included all the reconstructions that presented the same number of changes than the observed pattern (hypergeometric probabilities) were, in general, higher than those calculated when only MPRs were included (simulation analysis); however, both exhibited the same trend (Table 1). The correlation pattern was also observed when the analyses only included species with IRs of 11 bp (Table 2).

Stem length. None of the analyses yielded significant results for the correlation between the occurrence of inversions and the length of the stems (Table 3). When the threshold between long and short IRs was set to 11 bp and the sampling included all the species of each genera, there was a tendency to higher frequency of inversions on branches with long IRs. However, this pattern was non-significant (Table S3). In contrast to ΔG analysis, both the optimization of IR lengths from the terminals (either binary or continuous coding), and the analysis that derived IR lengths from ancestral hairpin structures gave very similar results. The probabilities calculated when the null distribution included all the reconstructions that present the same number of changes as the observed were very similar to those calculated when only MPRs were included in the null distribution. The different cost regimes and methods to optimize inversions (Binary coding or FSI) did not affect the results of the comparative analysis.

Discussion

A distinctive feature of the inversion analysed in the present study is its multiple occurrence in diverse lineages of Angiosperms (monocots, magnoliids, rosids). Most of the SI previously documented presented lower levels of recurrence and they generally involved closely related taxa (Kelchner and Wendel, 1996; Kelchner and Clark, 1997; Mes et al., 2000; Kim and Lee, 2005; Swangpol et al., 2007). The only SI with a widespread phylogenetic distribution and a similar level of recurrence was documented by Bain and Jansen (2006) in the *psb*A-*trn*H intergenic region. Nonetheless, these authors analysed the inversions by performing pairwise comparisons and not by optimizing them on a tree.

The SI analysed in the present study could be found throughout the angiosperms because of the high conservation level of the hairpin structure. This conservation is unexpected given the non-coding nature of this

region. Graham and Olmstead (2000) observed a similar situation in a SI found in early-derived lineages of angiosperms, and suggested that this conservation is associated with the role that this structure has in the stabilization of the corresponding transcripts. Accordingly, SI recurrence may be explained just as a byproduct of the hairpin function (Kim and Lee, 2005), without invoking any adaptive explanation.

Hairpin stability and inversion events

The present study suggests that the occurrence of the SI in *psb*C–*trn*S intergenic region is correlated with higher hairpin stability. This result is in accordance with Kelchner and Wendel's (1996) analysis of a different SI and points to an extension of the pattern described in that analysis in two different directions: First, the observation of similar patterns in SIs present in different intergenic regions suggests that this correlation is common to SIs situated throughout the chloroplast genome. Second, the pattern originally observed within Bambusoideae is now documented in angiosperms as a whole. The analyses of SIs situated in other regions of chloroplast genome with approaches similar to those employed here, will help to further test the generality of this pattern.

Kelchner and Wendel (1996) suggested that higher hairpin stability increases the chance for hairpin formation that in turn increases the frequency of SI. This idea implicitly considers that hairpin structure is involved in the process that finally leads to the inversion (hereafter hairpin hypothesis). In that analysis, the differences in hairpin stability among taxa were caused by changes in IR lengths. In contrast, we found no evidence that IR lengths affected the frequency of the inversion in *psb*C–*trn*S intergenic region. Changes in hairpin stability in the present analysis were mainly caused by mutations that produced imperfect pairing in the stems (quasipalindromes). Under the hairpin hypothesis, both observations can be explained in a simple way: short IRs and imperfect pairing decrease the chances of inversions by affecting hairpin stability (Fig. 6a).

Even if the correlation between hairpin stability and SIs can be perfectly explained in the context of the hairpin hypothesis, alternative explanations are also possible. For some of the molecular mechanisms proposed to explain SI occurrence (double-stranded slippage, Leach *in* Slupska et al., 2000; excision and religation of loop sequences, Kelchner and Wendel, 1996; intrastructure recombination, Kim and Lee, 2005; DNA-mediated strand invasion, Musumeci et al., 2000) the formation of the hairpin is a crucial step in this process. However, alternative models (no-hairpin models) do not consider that hairpins are required (repairing of an accidental double-break in palindromes, Slupska et al., 2000; intramolecular recombination over double-stranded DNA, Musumeci et al., 2000; intermolecular recombination, Graham and Olmstead, 2000). The action of intramolecular or intermolecular recombination in the DNA inversion process is a common characteristic of the no-hairpin proposals. It is well documented that there is a strong dependency of the recombination frequency on the length and similarity of the intervening sequences (Rubnitz and Subramani, 1984; Shen and Huang, 1986; Bazemore et al., 1997; Sagi et al., 2006; Barnes and McCulloch, 2007). Consequently, it is possible that IR lengths and imperfect pairings affect the frequency of inversions by altering the efficiency of the recombination step. Under this alternative explanation, the correlation between higher hairpin stability and inversion occurrence might not be explained by a cause–effect relationship between these variables (as is explained under the hairpin hypothesis), but just as a consequence that IR lengths and imperfect pairing also affect hairpin stability (Fig 6b).

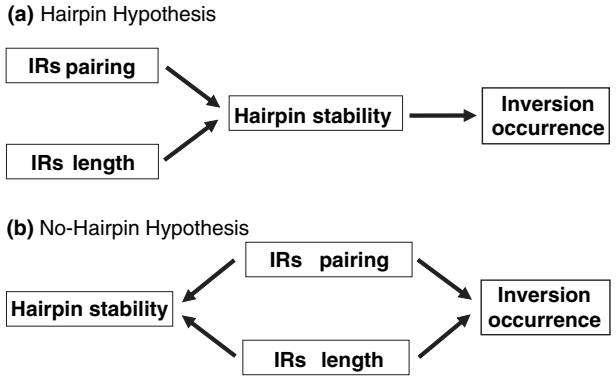

Fig. 6. Interpretation of the evidence about SCI evolution under two hypotheses proposed for inversion occurrence. In hairpin hypothesis (a) the stability of the hairpin would directly affect inversion occurrence. Pairing and IRs length indirectly affect inversion occurrence by affecting hairpin stability. In no-hairpin hypothesis (b) pairing and IRs length would directly and independently affect inversion occurrence.

In summary, we found clear evidence that supports a correlation pattern between hairpin stability and the occurrence of a SI situated in the *psb*C–*trn*S intergenic region. This result is in agreement with Kelchner and Wendel (1996) hypothesis that hairpin formation is involved in SI occurrence. Nevertheless, an alternative explanation where the frequency of inversions is independently affected by the IR lengths and imperfect pairing, without the mediation of hairpin formation, cannot be discarded. The formulation of more detailed predictions from the models, and the analysis of new SIs will contribute to a better comprehension about the particularities of these molecular events.

Homology and ancestral assignments

In the present study, different methods were evaluated to assign ΔG values to internal nodes. When the

ancestral values were obtained by optimizing ΔG values from the terminals, some of the most parsimonious reconstructions wrongly assigned high ΔG values to some deep nodes. This occurred because high ΔG values in different lineages were considered as derived from a common ancestor. However, when the sequences themselves were optimized using Direct Optimization or Iterative Pass Optimization, the ancestral sequences presented low ΔG values for those nodes, with high ΔG values being independently derived in several lineages by different substitutions in the IRs. The distinction of phenotypic states (ΔG values) that, although similar, were not derived from a common ancestor, was possible because the relationship between the genotype and the phenotype is very simple in this case: a substitution in the IRs produces a change in the corresponding ΔG values. This situation contrasts sharply with that of most of the morphological characters generally considered in phylogenetic analyses. Owing to the complex network of genetic interactions that defines these characters, the evaluation of the heritable changes that affect them is impossible in most of the cases, leaving similarity as the most reliable clue about common origin.

Given the heuristic nature of DO, these methods do not guarantee that the reconstructions are globally optimal. In addition, as currently implemented in POY IV Beta, only one group of optimal ancestral assignments can be obtained. This might cast doubts on the ancestral structure assignments and, concomitantly, on the phylogenetic comparison results. However, different lines of evidence stand against this possibility. First, the high conservation found in the IRs reduces the possibility of ambiguities in the ancestral sequence reconstructions. Second, the ancestral secondary structures were the same in all the analyses, irrespective of cost regimes or optimization algorithms. Third, even though it is possible that there were alternative optimal ancestral assignments than those analysed, this does not imply that ΔG values do in fact change. Finally, there is no reason to suppose that the reconstructions evaluated were biased in favour of the hypothesis tested.

Lineage effect

A criticism made of some comparative methods, CCT among them (Maddison, 2000) is that their outcome could be wrong when the changes in the response variable are clustered on the tree. Under these conditions, changes in the response variable may be caused by another factor that changes in the same branch as the predicting variable giving the false impression of a correlation between the variables analysed. In our analysis, the multiple origins of high ΔG values make unlikely this possibility as this third character should coincidentally change on all these branches. However, other factors unrelated (either directly or indirectly) with ΔG, can be conjectured to be responsible for this pattern. An alternative explanation may be related to changes in the enzymatic machinery that mediates the inversions: a change in this mechanism might alter the frequency of inversions. From this alternative explanation it is possible to derive a prediction: the same pattern should be observed in SIs situated in other regions of the plastome. Although not strictly tested, this prediction seems to have no empirical support. In *psb*C–*trn*S intergenic region, two large groups that presented high ΔG values lack inversions: the Poaceae family and Asterids. However, SIs are very common for these lineages in other chloroplast DNA regions: within Poaceae there are SIs reported for *Zea–Saccarum* and, *Oryza–Triticum* pairs (Kim and Lee, 2005) and in Bambusoideae (Kelchner and Wendel, 1996). Within Asterids, there are SIs documented between *Nicotiana* and *Atropa* (Kim and Lee, 2005) and between species of *Jasminum* and *Taraxacum* respectively (Mes et al., 2000). A common feature of all these cases is that the hairpins present high thermodynamic stability. This data would contradict the idea that changes in the enzyme machinery, as well as any other factor that affects the occurrence of SIs throughout the chloroplast genome, are responsible for the pattern observed in the present study.

Possible improvements on SI optimization

Different improvements are possible on the approach proposed here to optimize SI. In line with previous studies about SI, we assessed the occurrence of inversions by evaluating the sense of the spacer. However, most of the mechanism proposed to explain the inversions require that either part or the complete palindromes may also be inverted in a symmetric manner with respect to the spacer. If the palindromes are perfect, any inversion that involves part or the complete palindromes will only produce observable changes in the spacer but not in the inverted repeats. Hence, either the evaluation of the sense of the spacer or the evaluation of all the possible points of inversion will give the same results. When the palindromes are not perfect, an inversion that involves these quasipalindromes might not only change the spacer orientation, but may also modify the original flanking sequences. Even in this case, inversions may also be evaluated by analysing the sense of the spacer, although a more accurate procedure is to evaluate each of the possible limits of inversions and choose those limits that minimize the editing cost between sequences. This approach will allow dealing with sequences that present different limits between loop and stem in different species, as is the case when a mutation in the last base of the stem produces a mispairing. Exact and heuristic

solutions have already been developed to calculate editing costs when the limits of the inversions are not known *a priori* (Schöniger and Waterman, 1992; do Lago et al., 2005).

Another issue of SIs optimization is related to the heuristics. The method presented here is a modification of Fixed States Optimization, the rougher heuristic in the framework of dynamic homologies. This approach has the advantage that, once the Sankoff matrix is built, any software that deals with Sankoff characters (such as TNT) can be used to optimize inversions. However, better results are expected if methods such as Direct Optimization or Iterative Pass Optimization are modified in such a way to cope with inversions.[2]

Future prospects

We consider that future research on SI should follow three directions. First, as already indicated, studies such as those presented here should be performed in other SI to evaluate the generality of our results. Second, the results obtained in the phylogenetic analyses should be contrasted with experimental evidence, in particular with the aid of DNA recombination techniques. The combined assessment of both sources of evidence will help to have a better comprehension of the mechanism underling these molecular events. Finally, the utility of SIs to infer phylogenetic relationships should be further explored. Although in the earliest analyses SIs were considered as an unreliable source of phylogenetic information (Kelchner and Wendel, 1996; Sang et al., 1997), recent studies suggest that they are, at least for some groups, highly valuable (Bain and Jansen, 2006). We predict that the analysis of SI within a dynamic homology framework will result in a renewed interest in the use of these characters in plant phylogenetic studies.

Acknowledgements

We thank Pablo Goloboff and Norberto Giannini for helpful comments on the manuscript. We are indebted to Pablo Goloboff for his crucial help in TNT scripts programming. SAC is a fellow of Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET, Argentina). BOS and JCV are members of Carrera del Investigador Científico of CONICET. This study was supported by UBACYT_X321 (BOS), UBACTY_X201 and PIP-CONICET 5122 (JCV) grants.

[2]In November 2007, we were told by a POY IV author (Andrés Varón) that a similar approach to FSI would be implemented in this software. They also inform us that in the future it would be possible to analyse inversions by DO and IP.

References

Bain, J.F., Jansen, R.K., 2006. A chloroplast DNA hairpin structure provides useful phylogenetic data within tribe Senecioneae (Asteraceae). Can. J. Bot. 84, 862–868.

Barker, N.P., Clark, L.G., Davis, J.I., Duvall, M.R., Guala, G.F., Hsiao, G.F., Kellogg, E.A., Linder, H.P., 2001. Phylogeny and subfamilial classification of the grasses (Poaceae). Grass Phylogeny Working Group. Ann. Mo. Bot. Gard. 88, 373–457.

Barnes, R.L., McCulloch, R., 2007. *Trypanosoma brucei* homologous recombination is dependent on substrate length and homology, though displays a differential dependence on mismatch repair as substrate length decreases. Nucleic Acids Res. 35, 3478–3493.

Bazemore, L.R., Folta-Stogniew, E., Takahashi, M., Radding, C.M., 1997. RecA tests homology at both pairing and strand exchange. Proc. Natl Acad. Sci. USA 94, 11863–11868.

Blakeley, E., Rennie, K., Jones, L., Elstener, M., Chrzanowska-Lightowlers, Z., White, C., Shield, J., Pilz, D., Turnbull, D., Poulton, J., Taylor, R., 2006. Sporadic intragenic inversion of the mitochondrial DNA MTND1 gene causing fatal infantile lactic acidosis. Pediatr. Res. 59, 440–444.

Bremer, B., Bremer, K., Heidari, N., Erixon, P., Anderberg, A.A., Olmstead, R.G., Kållersjö, M., Barkhordarian, E., 2002. Phylogenetics of asterids based on 3 coding and 3 non-coding chloroplast DNA markers and the utility of non-coding DNA at higher taxonomic levels. Mol. Phylogenet. Evol. 24, 274–301.

Catalano, S.A., Vilardi, J.C., Tosto, D., Saidman, B.O., 2008. Molecular phylogeny and diversification history of *Prosopis* (Fabaceae: Mimosoideae). Biol. J. Linn. Soc. 93, 621–640.

Dumolin-Lapègue, S., Pemonge, M.-H., Petit, R.J., 1998. Association between chloroplast and mitochondrial lineages in oaks. Mol. Biol. Evol. 15, 1321–1331.

Farris, J., 1970. Methods for computing Wagner trees. Syst. Zool. 19, 83–92.

Fisher, R.A., 1935. The logic of inductive inference. J. R. Statist. Soc. A, 98, 39–54.

Goloboff, P.A., Farris, J.S., Nixon, K., 2003. TNT: Tree Analysis Using New Technology. Program and documentation available at http://www.zmuc.dk/public/phylogeny.

Goloboff, P.A., Mattoni, C.I., Quinteros, A.S., 2006. Continuous characters analyzed as such. Cladistics 22, 589–601.

Graham, S.W., Olmstead, R.G., 2000. Evolutionary significance of an unusual chloroplast DNA inversion found in two basal angiosperm lineages. Curr. Genet. 37, 183–188.

Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. 41, 95–98.

Harvey, P.H., Pagel, M., 1991. The Comparative Method in Evolutionary Biology. Oxford University Press, Oxford.

Jansen, R.K., Palmer, J.D., 1987. A chloroplast DNA inversion marks an ancient evolutionary split in the sunflower family (Asteraceae). Proc. Natl Acad. Sci. USA 84, 5818–5822.

Kelchner, S.A., Clark, L.G., 1997. Molecular Evolution and phylogenetic utility of the chloroplast rpl16 intron in *Chusquea* and the Bambusoideae (Poaceae). Mol. Phylogenet. Evol. 8, 385–397.

Kelchner, S.A., Wendel, J.F., 1996. Hairpins create minute inversions in non-coding regions of chloroplast DNA. Curr. Genet. 30, 259–262.

Kim, K.J., Lee, H.-L., 2005. Widespread occurrence of small inversions in the chloroplast genomes of land plants. Mol. Cells 19, 104–113.

do Lago, A.P., Muchnik, I., Kulikowski, C., 2005. A sparse dynamic programming algorithm for alignment with non-overlapping inversions. Theoret. Informatics Appl. 39, 175–189.

Lavin, M., Herendeen, P.S., Wojciechowski, M.F., 2005. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary. Syst. Biol. 54, 575–594.

Maddison, W.P., 1990. A method for testing the correlated evolution of two binary characters: are gains or losses concentrated on certain branches of a phylogenetic tree? Evolution 44, 539–557.
Maddison, W.P., 2000. Testing character correlation using pairwise comparisons on a phylogeny. J. Theor. Biol. 202, 195–204.
Markham, N.R., Zuker, M., 2005. DNAMelt web server for nucleic acid melting prediction. Nucleic Acids Res. 33, W577–W581.
Mes, T., Kuperus, P., Kirschner, J., Stepanek, J., Oosterveld, P., Storchova, H., den Nijs, J., 2000. Hairpins involving both inverted and direct repeats are associated with homoplasious indels in noncoding chloroplast DNA of *Taraxacum* (Lactuceae: Asteraceae). Genome 43, 634–641.
Milligan, B., Hampton, J., Palmer, J., 1989. Dispersed repeats and structural reorganization in subclover chloroplast DNA. Mol. Biol. Evol. 6, 355–368.
Moret, B.M.E., Siepel, A.C., Tang, J., Liu, T., 2002. Inversion medians outperform breakpoint medians in phylogeny reconstruction from gene-order data. In: Guigó, R. and Gusfield, D. (Eds.), 2nd Workshop on Algorithms in Bioinformatics – WABI 2002. Springer-Verlag, Berlin, pp. 521–536.
Musumeci, O., Andreu, A.L., Shanske, S., Bresolin, N., Comi, G.P., Rothstein, R., Schon, E.A., DiMauro, S., 2000. Intragenic inversion of mtDNA: a new type of pathogenic mutation in a patient with mitochondrial myopathy. Am. J. Hum. Genet. 66, 1900–1904.
Needleman, S.B., Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48, 443–453.
Olmstead, R.G., Sweere, J.A., Spangler, R.E., Bohs, L., Palmer, J.D., 1999. Phylogeny and provisional classification of the Solanaceae based on chloroplast DNA. In: Nee, M., Symon, D.E., Lester, R.N., Jessop, J.P. (Eds.), Phylogeny and Provisional Classification of the Solanaceae Based on Chloroplast DNA. Royal Botanic Gardens, Kew, pp. 111–137.
Palmer, J.D., 1985. Comparative organization of chloroplast genomes. Annu. Rev. Genet. 19, 325–354.
Raubeson, L., Jansen, R., 1992. Chloroplast DNA evidence on the ancient evolutionary split in vascular land plants. Science 255, 1697–1699.
Rubnitz, J., Subramani, S., 1984. The minimum amount of homology required for homologous recombination in mammalian cells. Mol. Cell. Biol. 4, 2253–2258.
Sagi, D., Tlusty, T., Stavans, J., . High fidelity of RecA-catalyzed recombination: a watchdog of genetic diversity. Nucleic Acids Res. 34, 5021–5031.
Sang, T., Crawford, D.J., Stuessy, T.F., 1997. Chloroplast DNA phylogeny, reticulate evolution, and biogeography of *Paeonia* (Paeoniaceae). Am. J. Bot. 84, 1120–1136.
Sankoff, D.D., Rousseau, P., 1975. Locating the vertices of a Steiner tree in arbitrary space. Math. Program. 9, 240–246.
Schöniger, M., Waterman, M.S., 1992. Local algorithm for DNA sequence alignment with inversions. Bull. Math. Biol. 54, 521–536.
Shen, P., Huang, H.V., 1986. Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. Genetics, 112, 441–457.
Slupska, M.M., Chiang, J.-H., Luther, W.M., Stewart, J.L., Amii, L., Conrad, A., Miller, J.H., 2000. Genes involved in the determination of the rate of inversions at short inverted repeats. Genes Cells 5, 425–437.
Stern, D.B., Gruissem, W., 1987. Control of plastid gene expression: 3′ inverted repeats act as mRNA processing and stabilizing elements, but do not terminate transcription. Cell 51, 1145–1157.
Stern, D.B., Jones, H., Gruissem, W., 1989. Function of plastid mRNA 3′ inverted repeats: RNA stabilization and gene-specific protein binding. J. Biol. Chem. 264, 18742–18750.
Swangpol, S., Volkaert, H., Sotto, R., Seelanan, T., 2007. Utility of selected non-coding chloroplast DNA sequences for lineage assessment of *Musa* interspecific hybrids. J. Biochem. Mol. Biol. 40, 577–587.
Varón, A., Vinh, L., Bomash, I., Wheeler, W., 2007. POY 4.0 Beta 2398. American Museum of Natural History. http://research.amnh.org/SIcomp/projects/poy.php.
Wenzel, J.W., Carpenter, J.M., 1994. Comparing methods: adaptive traits and tests of adaptation. In: Eggleton, P., Vane-Wright, R.I. (Eds.), Comparing Methods: Adaptive Traits and Tests of Adaptation. Academic Press, London, pp. 79–101
Wheeler, W., 1996. Optimization alignment: the end of multiple sequence alignment in phylogenetics? Cladistics 12, 1–9.
Wheeler, W., 1999. Fixed character states and the optimization of molecular sequence data. Cladistics 15, 379–385.
Wheeler, W., 2003. Iterative pass optimization of sequence data. Cladistics 19, 254–260.

Supporting Information

Additional Supporting Information may be found on online version of this article:

Table S1 Name and accession number of the species downloaded from GenBank. The column *Analysis* indicates whether the sequence was included in both analysis (2 species per genera and complete sampling; II) or only in the 2 per genera analysis (I). The sequences that lack the hairpin structure (N1), presented a strongly modified hairpin structure (N2), or had a different limits between loop and stems (N3) were excluded from the analyses. In ΔG column, free energy values associated with the corresponding hairpin structure are indicated. Motif = state assign to each species in binary coding of the inversion. IR = inverted repeats.

Table S2 Comparative analysis results. Relationship between free energy (ΔG) of the hairpin and inversion occurrence. The sampling includes all the species available for each genus. *Ancestral secondary structure analysis. Values calculated for the combination of topology-inversion reconstruction that most contradicts the hypothesis evaluated. Hyper, probabilities calculated using a cumulative hypergeometric distribution

Table S3 Comparative method results. Relationship between inverted repeats (IRs) length and inversion occurrence. The sampling includes all the species available for each genus. Hyper, probabilities calculated using a cumulative hypergeometric distribution

Table S4 Comparative method results. Relationship between free energy (ΔG) of the hairpin structure and inversion occurrence. The sampling includes all the species available for each genus that presented inverted repeats of 11 bp. *Ancestral secondary structure analysis. Values calculated for the combination of topology-inversion reconstruction that most contradicts the hypothesis evaluated. Hyper, probabilities calculated using a cumulative hypergeometric distribution

Table S5 Comparative method results. Relationship between free energy of the hairpin structure and inversion occurrence considering an alternative resolu-

tion among monocots, eudicots and magnoiids. The sampling involved a maximum of two species per genera. *Ancestral secondary structure analysis. Values calculated for the combination of topology-inversion reconstruction that most contradicts the hypothesis evaluated. Hyper, probabilities calculated using a cumulative hypergeometric distribution

Table S6 Comparative method results. Relationship between free energy of the hairpin structure and inversion occurrence when alternative penalty costs are evaluated: Indel, 10; Inversion, 3; substitution, 1. The sampling involved a maximum of two species per genera. *Ancestral secondary structure analysis. Values calculated for the combination of topology-inversion reconstruction that most contradicts the hypothesis evaluated. Hyper, probabilities calculated using a cumulative hypergeometric distribution

Fig. 1S. Histogram of free energy values.

Fig. 2S. Histogram of inverted repeats length values.

Fig. 3S. Condense topology used in the comparative analysis. Up to 50 resolutions of the polytomies presented in this cladogram were evaluated in each analysis.

Fig. 4S. Artificial dataset showing differences in the ancestral sequences assignments when the original sequences or the reverse complement were considered in the optimization.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Appendix 1

When defining cost regimes in Fixed State Optimization with inversions (FSI) some logical limits are implied. Costs should be given in such a way that if one sequence of length L is the perfect reverse complement of a second sequence, an inversion event should be accepted as optimal instead of L substitution events. One possible way to define a set of cost that meet this condition, is to consider inversion penalties that are lower than the lowest cost calculated between each sequence and its reverse complement. The sets of cost considered here met this condition.

As the Sankoff matrix of transformation costs between fragments is built, ties can occur if the editing cost between two sequences is equal to the editing cost between the first sequence and the inverted second sequence plus the inversion penalty cost. This ambiguity was taken into account when moving along the tree to define on which branches an inversion had occurred. Two different set of costs were considered in the analysis. The results obtained considering one of them are shown in the manuscript (Indel penalty (g) = 10, substitution penalty (s) = 1 and inversion penalty (i = 1). The results obtained with a second set of costs (g = 10, s = 1 and i = 3) are shown in the Supporting Information.

The FSI requires the limits of the spacer to be the same for all the sequences analysed. In fact, this is the case for most of the sequences considered in this analysis. Those sequences that did not present the same limits (approximately 10% of the total sequences) were excluded from the analysis.

The FSI was implemented with a script (supplementary data online) using the macro language of TNT. A problem with its implementation on TNT macro language is the current maximum number of states allowed, which is 31. Hence, only 31 different spacers sequences could be evaluated at once. As the datasets present 35 different spacers motifs, we decided to randomly exclude four terminals. Different groups of taxa were excluded in order to evaluate whether this can affect the final results. As the conclusions did not change, the results obtained considering one of the reduced matrices are shown.

Appendix 2

In CCT, changes in the response variable are sorted out in gains and losses. In the case of inversions, all the changes are equivalent. Hence, the null hypothesis in our analysis can be restated as (cf. Maddison, 1990): C changes in the first character, are randomly distributed on the tree regardless the value of the second character The probability of a particular distribution of inversions on the tree is equal to B/A, where A is the number of possible ways to have the observed number of inversions on the tree and B is the number of possible distributions that present the same number of inversions on branches with low and high free energy (or long/short IRs). In CCT analysis, where gains and losses are differentiated, this probability cannot be simply calculated by counting the branches, since it necessary to take into account the topology during the calculations (Maddison, 1990; Harvey and Pagel, 1991). This occurs because if the response character is lost in one branch it cannot be lost once again in a descendant branch. When dealing with inversions this is not the case: a sequence can be inverted in a branch and once again in a descendant branch. Consequently, once the number of branches presenting each state of ΔG is counted and the branches where inversions occurred is verified, the tree topology can be left aside and the probability can be calculated using a hypergeometric distribution. This distribution is the same used in the Fisher exact test (Fisher, 1935), Therefore, when gains and losses are not differentiated, CCT analysis is equivalent to perform a 2×2 contingency table with the branches being classified by: (1) inversion/no inversion, (2) high ΔG/low ΔG. In order to evaluate the null hypothesis (inversions randomly distributed without regard the value of ΔG or IR lengths) it is also necessary to compute the probability of any pattern of changes that is more extreme than the observed one. Hence, a hypergeometric cumulative distribution should be used.

When the probabilities are calculated in the way previously indicated, the null distribution includes all those reconstructions that present the same number of inversions irrespective of whether these are most parsimonious reconstructions (MPR) or not (Note: in its original description CCT includes both). Wenzel and Carpenter (1994) pointed out that only MPRs should be included in the null distribution. The hypergeometric distribution cannot be used to calculate this probability. Hence, in order to compare both probabilities, we performed a simulation analysis that only includes in the null distribution MPRs with the same number of changes than the observed pattern. The simulation was written in TNT macro (available upon request to the author).

In CCT and the comparative method considered here, branches are separated in two possible states (in our case low/high energy value). A branch with *low* state is a branch that presents both the subtending and the derived node with the state *low*. However, those branches were the predicting variable change cannot be computed as neither *low* nor *high*. Hence, these branches were not considered in our analysis. To evaluate the effect of different MPRs in the results, we considered up 50 randomly chosen reconstructions for each character (easily accomplished with the *iterrecs* option of TNT). Besides, in order to account for topological uncertainty, up to 50 different resolutions of the polytomies were evaluated.