Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Palevol 2 (2003) 241–250
COMPTES RENDUS PALEVOL

Évolution
(Mécanismes évolutifs, microévolution)

Independent evolution of *Cycloidea*-like sequences in several angiosperm taxa

Catherine Damerval [a,b,*], Michaël Manuel [c]

[a] *Station de génétique végétale, INRA/UPS/INA-PG/CNRS UMR 8120, La Ferme du Moulon, 91190 Gif-sur-Yvette, France*
[b] *Groupe « Développement et Évolution », université Pierre-et-Marie-Curie/CNRS UMR 7622, 9, quai Saint-Bernard, case 241, 75252 Paris cedex 05, France*
[c] *Service commun de biosystématique, université Pierre-et-Marie-Curie, 9, quai Saint-Bernard, case 241, 75252 Paris cedex 05, France*

Received 30 April 2002; accepted 3 March 2003

Presented by Philippe Taquet

Abstract

The TCP family of putative transcriptional factors, defined by the founding members *Teosinte branched1*, *Cycloidea* and *PCF*, is characterised by a specific basic helix-loop-helix domain. CYC and PCF subfamilies have been defined on the basis of sequence differences in conserved domains. Twenty-four coding sequences containing a TCP domain were found in the complete genome of *Arabidopsis thaliana* using BLAST searches. A neighbour-joining analysis of 112 sequences from various angiosperm taxa, belonging to the TCP family, suggests (*i*) homoplasy for the presence/absence of an R domain, (*ii*) multiple independent duplications leading to the wide diversity of CYC subfamily sequences in the five plant families sampled. ***To cite this article: C. Damewal, M. Manuel, C. R. Palevol 2 (2003).***

Résumé

Évolution indépendante de séquences homologues de *Cycloidea* dans plusieurs taxons d'angiospermes. La famille de facteurs de transcription «TCP », définie initialement à partir des gènes *Teosinte branched1*, *Cycloidea* et *PCF*, est caractérisée par un domaine hélice-boucle-hélice basique original. Les sous-familles CYC et PCF ont été définies d'après divers domaines conservés. Vingt-quatre séquences codantes comportant un domaine TCP ont été trouvées dans le génome complet d'*Arabidopsis thaliana*. Une analyse de *neighbour-joining* sur 112 séquences de la famille TCP suggère (*i*) l'existence d'homoplasie pour la présence du domaine R, (*ii*) des événements multiples de duplication de gènes conduisant à une grande diversité de la sous-famille CYC, indépendants dans les cinq familles de plantes étudiées. ***Pour citer cet article : C. Damewal, M. Manuel, C. R. Palevol 2 (2003).***

* Corresponding author.
E-mail address: damerval@moulon.inra.fr (C. Damerval).

DOI: 10.1016/S1631-0683(03)00031-9

Keywords: TCP family; CYC subfamily; *Cycloidea*; *Arabidopsis*; Angiosperm; Molecular phylogeny; Evolution

Mots clés : Famille TCP ; Sous-famille CYC ; *Cycloidea* ; *Arabidopsis* ; Angiospermes ; Phylogénie moléculaire ; Évolution

Version française abrégée

Introduction

Les gènes *Cycloidea (CYC*) et *Dichotoma (DICH*) interviennent dans le contrôle du développement floral chez *Antirrhinum majus*, où ils sont responsables de la zygomorphie de la fleur adulte [16, 17]. Ces deux gènes codent des protéines comportant un domaine conservé de type bHLH (hélice-boucle-hélice basique) dénommé domaine TCP (du nom des gènes qui définissent la famille, *Teosinte Branched1* (*TB1*), *CYC* et *PCF*), qui caractérise une famille d'activateurs de transcription supposés [6]. Cette famille n'a jusqu'ici été trouvée que chez les végétaux. Les quelques membres dont on a étudié l'expression (*CYC*, *DICH*, *LCYC*-orthologue de *CYC* chez *Linaria vulgaris* [5], *TCP1*, *TCP2* et *TCP3* chez *Arabidopsis* [5–7, 16, 17]) semblent intervenir dans des processus de croissance et de développement (voir aussi [8, 9, 12].

Deux sous-familles sont définies sur la base de la séquence du domaine TCP et d'autres régions conservées [6]. La première, la sous-famille CYC, possède un signal de localisation nucléaire en deux parties, altéré dans la seconde, la sous-famille PCF. Des différences existent également dans la composition des hélices et de la boucle qui les sépare, ainsi que dans la longueur de la seconde hélice. Certains membres de la sous-famille CYC, dont CYC, DICH et TB1, comportent en outre un domaine riche en acides aminés polaires, situé à des distances variables en aval du domaine TCP, dit domaine R. Les membres de la sous-famille PCF comportent d'autres domaines conservés adjacents au domaine TCP [6].

La diversité des gènes homologues de *CYC* a été analysée dans deux familles botaniques de l'ordre des Lamiales *sensu* APG [2], les Veronicaceae *sensu* Olmstead et al. [18] et les Gesneriaceae [3, 20]. Une famille oligogénique de cinq membres au moins a été mise en évidence chez les Veronicaceae, et de deux membres chez les Gesneriaceae. Les duplications/pertes de gènes à l'origine de cette complexité semblent s'être produites de manière indépendante dans les deux taxons [3].

Afin de préciser la complexité de la famille TCP et l'évolution des gènes de la sous-famille CYC, nous avons tout d'abord recensé les séquences codantes comportant un domaine TCP dans le génome complet d'*Arabidopsis thaliana* (Brassicaceae), puis réalisé une phylogénie de 112 séquences extraites de Genbank, en axant l'analyse sur les gènes de la sous-famille CYC, principalement chez les eudicotylédones.

Matériel et méthodes

La recherche de séquences codantes comportant un domaine TCP dans le génome complet d'*Arabidopsis thaliana* a été faite par BLAST (BLASTp et tBLASTn, [1]) à l'aide de séquences complètes d'homologues de *CYC* et des domaines TCP de *CYC* et de *PCF1*.

Les séquences des gènes des sous-familles CYC et PCF ont été obtenues à partir de Genbank [3, 5, 15–17, 20]. L'échantillon de gènes *TB1* a été volontairement restreint après une analyse préliminaire, afin de focaliser l'étude sur les eudicotylédones.

Au total 112 séquences ont été analysées. L'alignement nucléotidique a été réalisé grâce au logiciel BioEdit (version 4.8.6) [13] et amélioré grâce aux séquences protéiques. La divergence importante des séquences en dehors des domaines TCP et R nous a conduits à ne conserver que ces deux domaines et la région qui les sépare pour les analyses phylogénétiques (834 positions). Pour les séquences dépourvues de domaine R (toutes celles de la sous-famille PCF et certaines de la sous-famille CYC), seule la séquence du domaine TCP a été prise en compte (177 nucléotides), les autres positions étant codées en données manquantes. Les délétions insérées pour maximiser l'alignement ont également été codées en données manquantes. Les analyses phylogénétiques ont été réalisées avec le logiciel PAUP (version 4.06b) [19] sur les alignements nucléotidiques et protéiques. La méthode du *neighbour-joining* (correction de Jukes et Cantor pour la distance nucléotidique [14]) a été appliquée à la matrice complète, et une analyse de parcimonie a été faite sur un échantillon restreint à 21 séquences.

Résultats et discussion

Vingt-quatre séquences codantes comportant un domaine TCP ont été trouvées dans le génome complet d'*Arabidopsis thaliana* (voir aussi [4]). Onze appartiennent à la sous-famille CYC (dont cinq comportent également un domaine R) et 13 à la sous-famille PCF (Tableau 1).

Une analyse par *neighbour-joining* a été réalisée sur ces 24 séquences et 88 séquences extraites de Genbank, représentant la diversité des séquences TCP connues jusqu'ici principalement chez les eudicotylédones. Les résultats obtenus sur les alignements nucléotidiques (834 positions) et protéiques (278 positions) sont comparables et seul l'arbre issu de l'analyse nucléotidique est présenté (Fig. 1). Comme nous souhaitions focaliser l'analyse sur l'évolution des gènes de la sous-famille CYC, nous avons arbitrairement raciné l'arbre sur la sous-famille PCF.

Les séquences dépourvues de domaine R ne forment pas un groupe monophylétique, mais apparaissent en trois groupes, qui comportent dans certains cas des séquences portant le domaine R. Il semble donc exister de l'homoplasie pour la présence/absence de ce domaine. L'hypothèse la plus parcimonieuse serait qu'il est apparu une fois à la base de la sous-famille CYC, et a été perdu plusieurs fois indépendamment (au minimum trois fois d'après l'arbre obtenu).

Par ailleurs, l'ensemble des séquences de Veronicaceae et de Gesneriaceae constitue un groupe monophylétique (*bootstrap* 81%). À l'intérieur de ce groupe, des groupes d'orthologie bien soutenus correspondent aux différents membres des familles oligogéniques mises en évidence chez ces deux familles botaniques par les études antérieures [3, 20]. Il est remarquable qu'aucun panachage de séquences issues des deux taxons n'apparaisse. De même, les duplications à l'origine des séquences de *Mimulus* (Phrymaceae) ont eu lieu indépendamment de celles des Veronicaceae et Gesneriaceae. Enfin, parmi les cinq séquences d'*Arabidopsis* possédant à la fois les domaines TCP et R, aucun orthologue, ni du gène *CYC* d'*A. majus*, ni d'aucun autre gène de l'analyse, n'apparaît. Nous confirmons, donc sur un échantillonnage taxinomique plus important que celui de Citerne *et al*. [3], que l'évolution des gènes de la sous-famille CYC s'est faite par des duplications/pertes multiples de gènes, qui ont eu lieu de manière indépendante dans les différents taxons. L'analyse de parcimonie confirme ces résultats (Fig. 2).

Cette évolution indépendante selon les taxons pose la question de la conservation de la fonction et du rôle des gènes de la sous-famille CYC. Cette fonction semble liée à la croissance et au développement ; la symétrie/asymétrie du territoire d'expression de ces gènes pourrait déterminer l'apparition de structures morphologiques symétriques ou non. Un échantillonnage taxinomique plus large ainsi que des études d'expression plus nombreuses sont indispensables pour mieux comprendre l'évolution de la sous-famille CYC, et pour examiner son rôle dans l'élaboration de structures morphologiques symétriques ou non.

1. Introduction

The TCP family of putative transcription factors is characterised by an original conserved basic-Helix-Loop-Helix (bHLH) domain that was initially found in Teosinte branched1 (TB1) of *Zea mays*, Cycloidea (CYC) and Dichotoma (DICH) from *Antirrhinum majus*, and two *Oryza sativa* DNA-binding proteins, PCF1 and PCF2. The initials of the founding members (TB1, CYC and PCF) were used to name the family [6]. So far, the TCP family appears to be restricted to plants.

Two subfamilies were defined based on the characteristics of the TCP domain and other conserved regions [6]. The first one (CYC subfamily), including CYC and DICH from *A. majus* and TB1 from *Z. mays*, has a bipartite nuclear localisation signal (NLS), while the second one (PCF subfamily), including the PCFs from *O. sativa*, has only a portion of the bipartite NLS. Moreover, differences in residue composition exist in the hydrophobic faces of the helices and the loop, and helix II is shorter in the PCF subfamily. Some members of the CYC subfamily are characterised by a second domain rich in polar residues, located at various distances downstream to the TCP domain, called the R domain [6]. A third conserved domain, rich in proline and serine residues, has been described as specific to the TB1 protein in Andropogoneae and a few other grasses [15]. Members of the PCF subfamily share conserved regions adjacent to the TCP domain [6].

Genes belonging to this family have been suggested to play a role in plant growth and development. *Cycloidea* (*CYC*) and *Dichotoma* (*DICH*) genes are in-

volved in the control of floral development in *Antirrhinum majus*, promoting a zygomorphic adult flower in wild-type genotypes [16, 17]. Both are expressed asymmetrically, in the dorsal part of developing floral meristems. *CYC* has been found to repress a D-cyclin gene in the dorsal stamen of *Antirrhinum* flower [12]. In maize, *TB1* is involved in arresting axillary meristem growth, reducing internode elongation in axillary branches and arresting petal and stamen development in female flowers [8, 9]. PCF1 and PCF2 were isolated on the basis of their ability to bind to promoter elements of a rice gene whose product is involved in DNA replication and cell cycle control. Among new genes discovered in this family [6], expression of *TCP1*, *TCP2* and *TCP3* genes of *Arabidopsis* are observed in developing floral meristems. Expression of *TCP2* and *TCP3* does not differ between dorsal and ventral part of the primordia [6], while expression of *TCP1* is transient in the dorsal part of the primordia [7].

Within the CYC subfamily, the diversity of so-called *Cycloidea*-like genes has been analysed in two angiosperm families belonging to the order Lamiales *sensu* APG [2]. In twelve species of Veronicaceae *sensu* Olmstead et al. [18], an oligogenic family of at least five members, called *CYC1A* and *1B*, *CYC2*, *CYC3* and *CYC4*, was found. The orthologues of four of the five genes were found in all sampled taxa, and *CYC3* was present in *Misopates orontium*, but not deeply investigated in the other taxa. *CYC1A* and *1B* are very similar and must have originated through a recent duplication [20]. In the family Gesneriaceae, several *Cycloidea*-like sequences, termed *GCYC*, were detected from most species studied [3]. Orthologues of *GCYC1* occurred in all studied taxa, with a recent duplication in the *Streptocarpus/Saintpaulia* clade giving birth to *GCYC1A* and *1B*. Orthologues of *GCYC2* were detected only in *Ramonda*, *Conandron* and *Haberlea*.

In the Lamiales *sensu* APG, a zygomorphic flower is predominant and believed to be the ancestral state [11]. Under this assumption, the actinomorphic flowers observed in many species among the Gesneriaceae should be derived. However, no clear relationship appeared between the number of *GCYC* genes or their sequence and flower actinomorphy or zygomorphy. Based on an average mutation rate estimated for the *CYC* gene in Veronicaceae (1.5×10^{-9} nucleotide per year), the divergence between (*CYC1*, *CYC2)* on the one hand, and (*CYC3*, *CYC4*) on the other hand was estimated to occur about 75 Myr ago, which suggests that orthologues should exist in other angiosperm families [20]. However, this molecular datation is strongly questioned by the fact that no orthologues of *CYC*, *DICH* and *LCYC* (an orthologue of *CYC* in *Linaria vulgaris* [5]) from Veronicaceae seem to exist among the *GCYC* sequences of Gesneriaceae [3].

The availability of the full genomic sequence of *Arabidopsis thaliana* (Brassicaceae) gives the opportunity to count and compare all members of the TCP family in a given genome. We included the 24 *Arabidopsis* sequences with a TCP domain also found by Cubas [4] in a phylogenetic analysis of *CYC*-like sequences available in Genbank, with a specific focus on eudicotyledon species. Our objectives were to investigate the evolution of the CYC subfamily, and to search for putative orthologues of *CYC*, which would give clues on possible involvement of *CYC*-like genes in floral symmetry in taxa other than Veronicaeae.

2. Material and methods

2.1. Search for TCP homologous coding sequences in the full genomic sequence of Arabidopsis

In a first step, translated sequences of Y16313 (*Antirrhinum majus*), AF208338 (*Sreptocarpus holstii*), AF208318 (*Ramonda myconi*), AF208335 (*Streptocarpus dunnii*), AF146880 and AF146873 (*Misopates orontium*), AF199465 (*DICH* from *Antirrhinum majus)*, AF161252 (*LCYC* from *Linaria vulgaris*) and AF146862 (*Linaria triornithophora*) were used to perform BLASTp [1] searches on the non-redundant database with an advanced search restricted to *Arabidopsis thaliana* (ORGN) and a threshold of 10^{-4} for the *e* value. To check that no sequence was missed using this procedure, (*i*) tBLASTn searches were performed on the *Arabidopsis* genome, using the nucleotide sequence of the *CYC* (Y16313) and the *PCF1* TCP domains, and (*ii*) every retrieved sequence was used to perform BLASTp on the full *Arabidopsis* genome.

2.2. CYC subfamily sequences

Eighty eight sequences were retrieved from Genbank (see Fig. 1): 48 from [20], 26 (the *Sinningia*

peloric mutant sequence being excluded) from [3], four corresponding to two *Mimulus* species, the *LCYC* gene of *Linaria vulgaris* [5], the originally isolated *CYC* and *DICH* sequences [16, 17], a *CYCLOIDEA*-like sequence from *Lycopersicon esculentum*, a *TCP3* homologous sequence from *Oryza sativa*, one *TB1* sequence from *Oryza sativa* and one *TB1* gene from *Populus*.

In a first step, all *TB1* sequences from the Andropogoneae analysed in [15] were retrieved and included in the neighbour-joining analysis (see below). However, since we wanted to focus on the eudicotyledon sequences, we present our results including only the *TB1* sequence from *Capillipedium parviflorum*, chosen on the basis of its position in the complete analysis.

2.3. Multiple sequence comparison and sequence analysis

DNA alignments were constructed using ClustalW as implemented in BioEdit version 4.8.6 [13], and visually refined on the basis of amino acid sequences. Because the sequences from Veronicaceae, Gesneriaceae, and *Mimulus*, all belonging to the Lamiales *sensu* APG, were very divergent outside the conserved domains, we chose to work in a region extending from the TCP to the R domains. The more variable interdomain region (i.e. between TCP and R domains) was kept because assessment of primary homology among the sequences appeared reasonable, at least among groups of related sequences. For TCP sequences devoid of an R domain, only the TCP domain was taken into account (177 nucleotides, 59 amino acid residues) and all other positions were coded as missing data. The total number of nucleic acid positions in the alignment was 834, of which 415 were constant. Phylogenetic analyses were performed using PAUP version 4.06b [19], from both amino acid and nucleotide matrices. The neighbour-joining (NJ) method was used with distance *p* for amino acid data and correction of Jukes and Cantor [14] for nucleotide data. Bootstrap was performed with 1000 replicates.

The choice of NJ can be criticized, because distance methods are not phylogenetic in their principle, and because NJ is less performing than other available methods, like maximum parsimony or maximum likelihood. However, in practice, NJ analyses efficiently recover the main sequence clusters from datasets such as ours (i.e. with many sequences and few informative positions), provided that significant nodes are distinguished from artefactual ones in the entirely resolved tree produced. Branch length and bootstrap values are the two criteria we used to identify significant nodes in the NJ trees. Maximum parsimony was also performed, but with the complete dataset, the high number of sequences relatively to the number of informative positions resulted in a huge number of equally parsimonious minimal topologies (as a result of the high degree of irresolution) and, for that reason, even the heuristic analysis could not be completed. A common practice in such a case is to show a strict-consensus tree computed from an arbitrary number of minimal trees, after the search has been aborted; but this results in neglecting a number of alternative minimal topologies, thus underestimating the irresolution. We preferred, as an alternative strategy, to reduce the number of sequences, on the basis of the initial NJ topology. A maximum parsimony analysis was thus performed on a subset of 21 sequences with a heuristic search strategy. Characters were unordered and were given equal weight, and gaps were treated as missing data. Random addition sequence of taxa was used with 100 replicates, followed by TBR (tree bisection-reconnection) branch swapping on the best trees. The option 'MULTREES on' was used. Bootstrap analyses (1000 replicates) were performed with the search option set to heuristic and random addition of taxa generating 50 starting tree replicates.

3. Results and discussion

3.1. TCP family sequences in Arabidopsis *genome*

The TCP domain and several *Cycloidea* protein sequences were used for searching sequences showing similarity in the complete genome of *Arabidopsis thaliana*. Twenty-four distinct sequences were found dispersed on the five chromosomes. It was not possible to establish the correspondence between all the sequences we retrieved and the accession numbers for the 24 sequences reported by Cubas [4]. Our designation of the TCPs is thus partly different from hers (Table 1). Thirteen sequences belong to the PCF subfamily. Among the eleven sequences belonging to the CYC subfamily, five have an R domain that is characteristic of CYC- or TB1-like proteins (*TCP1, TCP2,*

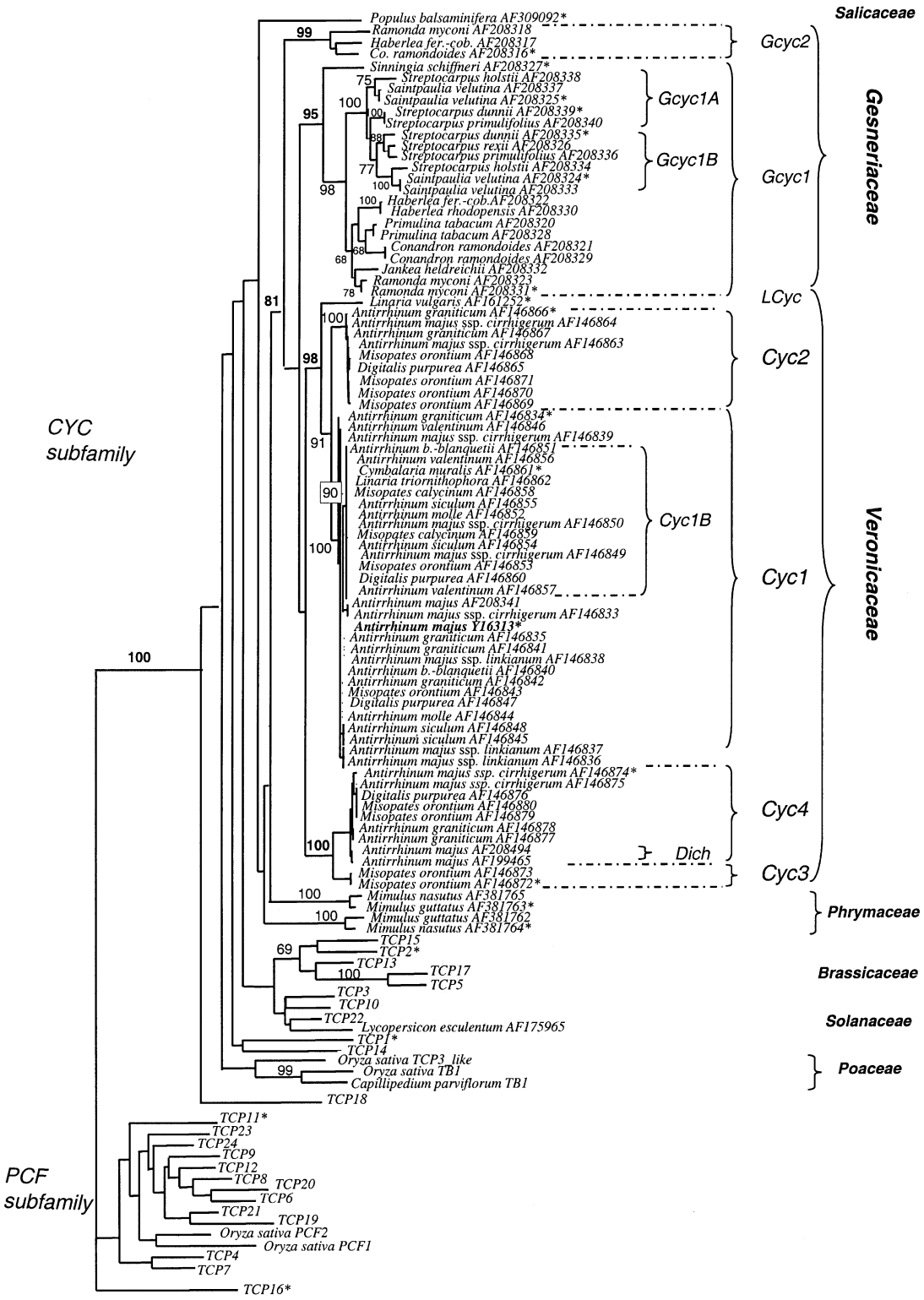
Salicaceae
Populus balsaminifera AF309092*
Ramonda myconi AF208318
Haberlea fer.-cob. AF208317
Co. ramondoides AF208316*
Gcyc2
Sinningia schiffneri AF208327*
Streptocarpus holstii AF208338
Saintpaulia velutina AF208337
Saintpaulia velutina AF208325*
Streptocarpus dunnii AF208339*
Streptocarpus primulifolius AF208340
Gcyc1A
Streptocarpus dunnii AF208335*
Streptocarpus rexii AF208326
Streptocarpus primulifolius AF208336
Streptocarpus holstii AF208334
Saintpaulia velutina AF208324*
Saintpaulia velutina AF208333
Gcyc1B
Gcyc1
Gesneriaceae
Haberlea fer.-cob.AF208322
Haberlea rhodopensis AF208330
Primulina tabacum AF208320
Primulina tabacum AF208328
Conandron ramondoides AF208321
Conandron ramondoides AF208329
Jankea heldreichii AF208332
Ramonda myconi AF208323
Ramonda myconi AF208331*
Linaria vulgaris AF161252*
LCyc
Antirrhinum graniticum AF146866*
Antirrhinum majus ssp. cirrhigerum AF146864
Antirrhinum graniticum AF146867
Antirrhinum majus ssp. cirrhigerum AF146863
Misopates orontium AF146868
Digitalis purpurea AF146865
Misopates orontium AF146871
Misopates orontium AF146870
Misopates orontium AF146869
Cyc2
CYC subfamily
Antirrhinum graniticum AF146834*
Antirrhinum valentinum AF146846
Antirrhinum majus ssp. cirrhigerum AF146839
Antirrhinum b.-blanquetii AF146851
Antirrhinum valentinum AF146856
Cymbalaria muralis AF146861*
Linaria triornithophora AF146862
Misopates calycinum AF146858
Antirrhinum siculum AF146855
Antirrhinum molle AF146852
Antirrhinum majus ssp. cirrhigerum AF146850
Misopates calycinum AF146859
Antirrhinum siculum AF146854
Antirrhinum majus ssp. cirrhigerum AF146849
Misopates orontium AF146853
Digitalis purpurea AF146860
Antirrhinum valentinum AF146857
Cyc1B
Antirrhinum majus AF208341
Antirrhinum majus ssp. cirrhigerum AF146833
Antirrhinum majus Y16313*
Antirrhinum graniticum AF146835
Antirrhinum graniticum AF146841
Antirrhinum majus ssp. linkianum AF146838
Antirrhinum b.-blanquetii AF146840
Antirrhinum graniticum AF146842
Misopates orontium AF146843
Digitalis purpurea AF146847
Antirrhinum molle AF146844
Antirrhinum siculum AF146848
Antirrhinum siculum AF146845
Antirrhinum majus ssp. linkianum AF146837
Antirrhinum majus ssp. linkianum AF146836
Cyc1
Veronicaceae
Antirrhinum majus ssp. cirrhigerum AF146874*
Antirrhinum majus ssp. cirrhigerum AF146875
Digitalis purpurea AF146876
Misopates orontium AF146880
Misopates orontium AF146879
Antirrhinum graniticum AF146878
Antirrhinum graniticum AF146877
Antirrhinum majus AF208494
Antirrhinum majus AF199465
Dich
Cyc4
Misopates orontium AF146873
Misopates orontium AF146872*
Cyc3
Mimulus nasutus AF381765
Mimulus guttatus AF381763*
Mimulus guttatus AF381762
Mimulus nasutus AF381764*
Phrymaceae
TCP15
TCP2*
TCP13
TCP17
TCP5
Brassicaceae
TCP3
TCP10
TCP22
Lycopersicon esculentum AF175965
Solanaceae
TCP1*
TCP14
Oryza sativa TCP3_like
Oryza sativa TB1
Capillipedium parviflorum TB1
Poaceae
TCP18
PCF subfamily
TCP11*
TCP23
TCP24
TCP9
TCP12
TCP8
TCP20
TCP6
TCP21
TCP19
Oryza sativa PCF2
Oryza sativa PCF1
TCP4
TCP7
TCP16*

Table 1
Name, characteristics and chromosomal assignation of TCP homologous coding sequences found in the *Arabidopsis* genome. The asterisk points to sequences numbered as in Cubas [4].
Tableau 1
Nom, caractéristiques et localisation chromosomique des séquences codantes comportant un domaine TCP, trouvées dans le génome complet d'*Arabidopsis*. L'astérisque indique les séquences qui ont pu être numérotées de manière identique à celle adoptée par Cubas [4].

Accession number/Protein_id	TCP designation	Domain types	Chromosome
AC002130/AAG00255	TCP1 [5]	CYC-type TCP, R	I
AC011914/AAG52043	TCP14	CYC-type TCP, R	I
AC073506/AAG50562	TCP15	CYC-type TCP, R	I
AP001303/BAB02213	TCP18*	CYC-type TCP, R	III
AL021710/CAA16719	TCP2 [5]	CYC-type TCP, R	IV
AF072134/ AAC24010	TCP3 [5]	CYC-type TCP	I
AC005311/AAC63845	TCP10*	CYC-type TCP	II
AP000370/BAA97066	TCP22	CYC-type TCP	III
AC011664/AAF14823	TCP13*	CYC-type TCP	III
AL357612/CAB93708	TCP17*	CYC-type TCP	V
AB008269/BAB10646	TCP5 [5]	CYC-type TCP	V
AC069273/AAG51130	TCP12	PCF-type TCP	I
AC013289/AAG52540	TCP24	PCF-type TCP	I
AC079131/AAG50759	TCP8 [5]	PCF-type TCP	I
AC007887/AAF79358	TCP23*	PCF-type TCP	I
AC006922/AAD31585	TCP11*	PCF-type TCP	II
AC003680/AAC06168	TCP21	PCF-type TCP	II
AL138649/CAB72153	TCP16*	PCF-type TCP	III
AY058874/AAL24261	TCP9 [5]	PCF-type TCP	III
AB026649/BAB01082	TCP20*	PCF-type TCP	III
AB025623/BAA97226	TCP19*	PCF-type TCP	V
AB007648/BAB11183	TCP4	PCF-type TCP	V
AL392174/CAC08333	TCP7 [5]	PCF-type TCP	V
AB010072/BAB09705	TCP6 [5]	PCF-type TCP	V

TCP14, *TCP15* and *TCP18*- see Table 1); they are distributed on three different chromosomes.

3.2. Diversity and evolution of sequences belonging to the CYC subfamily

After searches for TCP family sequences in GenBank, a dataset including 112 distinct members was constituted, including the 24 sequences from *A. thaliana*. A neighbour-joining distance analysis was performed from the 834-nucleotide alignments, including the TCP domain, the interdomain variable sequence, and the R domain. Since we were mostly interested in the evolution of the CYC subfamily in eudicotyledon species, the resulting tree has been arbi-

Fig. 1. Neighbour-joining tree resulting from the analysis of the 834-nucleotide alignment, including 112 genes, arbitrarily rooted on the PCF subfamily of TCP sequences. Bootstrap values above 60% are indicated on the branches (a few values above 60% on terminal branches are not indicated for clarity reasons). The 24 TCP coding sequences found in the *A. thaliana* genome are named TCP#, according to Table 1. Among the CYC subfamily, sequences devoid of an R domain appear in grey rectangles. The original *Cycloidea* gene of *A. majus* is in bold. Accession numbers are given following species names. Sequences used in the maximum parsimony analysis are indicated by a star.
Fig. 1. Arbre obtenu par la méthode du *neighbour-joining* sur la matrice nucléotidique de 834 positions et 112 gènes, avec les séquences de la sous-famille PCF en groupe externe. Les valeurs de *bootstrap* supérieures à 60% sont indiquées (quelques valeurs supérieures à 60% sur les branches terminales n'ont pas été indiquées pour des raisons de clarté). Les 24 séquences TCP d'*Arabidopsis* sont nommées TCP#, selon le Tableau 1. Les séquences de la sous-famille CYC sans domaine R figurent dans un rectangle gris ; le gène *Cycloidea*, originellement cloné chez *A. majus*, est indiqué en gras. Les numéros d'accession suivent les noms d'espèces. Les séquences utilisées dans l'analyse de parcimonie sont indiquées par un astérisque.

trarily rooted with the PCF subfamily as the outgroup (Fig. 1), but it should be recalled in the following that this tree is actually unrooted.

3.2.1. Evolution of the R domain

The eight genes from the CYC subfamily without an R domain (including the six genes from *Arabidopsis*) do not constitute a monophyletic group, but rather occur on three separate branches (Fig. 1). Some of them are closer to TCPs with an R domain, suggesting homoplasy for the presence/absence of the domain. Lack of resolution makes it difficult to appreciate the precise number of evolutionary events; furthermore, it is not possible to polarise the changes in the absence of a root. Assuming that the root is somewhere on the branch linking CYC and PCF subfamilies, the most parsimonious scenario would be an acquisition of the R domain in an ancestor of the whole CYC subfamily, followed by at least three independent losses. In any way, the occurrence of homoplasy for the presence/absence of the R domain should be borne in mind when assessing relationships of orthology and possible functional conservation of genes.

3.2.2. Relations of orthology and multiple independent duplications within the CYC subfamily

All sequences from Veronicaceae and Gesneriaceae constitute a monophyletic group (bootstrap value 81%), but among this clade, sequences from Veronicaceae or Gesneriaceae are not mixed. Instead, four well-supported clades can be recognised: *GCYC1* (bootstrap value 95%), *GCYC2* (bootstrap value 99%), (*CYC1*, *CYC2)* (bootstrap value 98%) and (*CYC3*, *CYC4)* (bootstrap value 100%), of which the first two contain only sequences from Gesneriaceae, and the last two contain only sequences from Veronicaceae. The *CYC* gene isolated for its major effect on floral symmetry in *Antirrhinum* belongs to the *CYC1* group of orthology, as well as the *LCYC* gene whose mutation is responsible for the *Linaria* peloric mutant [5]. *DICH*, also involved in zygomorphy in *Antirrhinum*, appears as a member of the *CYC4* group of orthology. No relation of orthology emerges between genes from Veronicaceae and Gesneriaceae, which indicates that in these two closely related families, gene diversity among the CYC subfamily was acquired independently, as already noticed by Citerne et al. [3]. In the same way, the four *Mimulus* sequences group in two well-supported clades that do not aggregate with sequences from either Veronicaceae or Gesneriaceae.

The *TCP1* gene from *A. thaliana* was suggested to be the orthologue of the *CYCLOIDEA* gene from *A. majus* [4, 7]. From our analysis, it is clear that *TCP1* is not an orthologue of *CYC*. Indeed, no relation of orthology can be proposed between particular TCP genes from *Arabidopsis* and Veronicaceae, respectively.

TB1 genes from *Oryza sativa* and *Capillipedium* form together a strongly supported clade. When all *TB1* sequences from ref [15] were included in the neighbour-joining analysis, orthology relationship could be well established among the genes of the grass species analysed, but no orthologous relationship occurred with genes from other taxa. In order to further examine relationships among the genes belonging to the CYC subfamily in eudicotyledon species, 21 sequences were selected from the main clades present in the neighbour-joining tree, including two PCF subfamily sequences (see Fig. 1) and were analysed using the maximum parsimony method. The dataset contained 264 informative characters. A single most parsimonious tree (1048 steps in length) was obtained (CI = 0.70, HI = 0.30, RI = 0.61, RC = 0.43), which, consistently with the neighbour-joining analysis on the full set of sequences, did not support any orthology relationship between genes from distinct taxonomic groups (Fig. 2).

3.3. Conclusion

The present study encompasses a broader gene sampling than in any previously published analysis of TCP-domain containing genes. The evolution of CYC subfamily genes appears to have occurred through multiple duplications (and possibly gene losses) taking place after the divergence between the various sampled taxonomic groups. The other possibility of multiple ancestral sequences homogenised through gene conversion, then following divergent evolution in the various taxa seems to be dismissed by the mapping of TCP family genes on different chromosomes in *Arabidopsis thaliana*.

The question of a possible conservation of function and expression across the different gene subgroups and taxa is presently unanswered. Expression data are recorded for six genes (*CYC*, *DICH*, *LCYC*, *TCP1*, *TCP2* and *TCP3*), and indicate a dorso-ventral asymmetric expression in developing meristem for *CYC*,

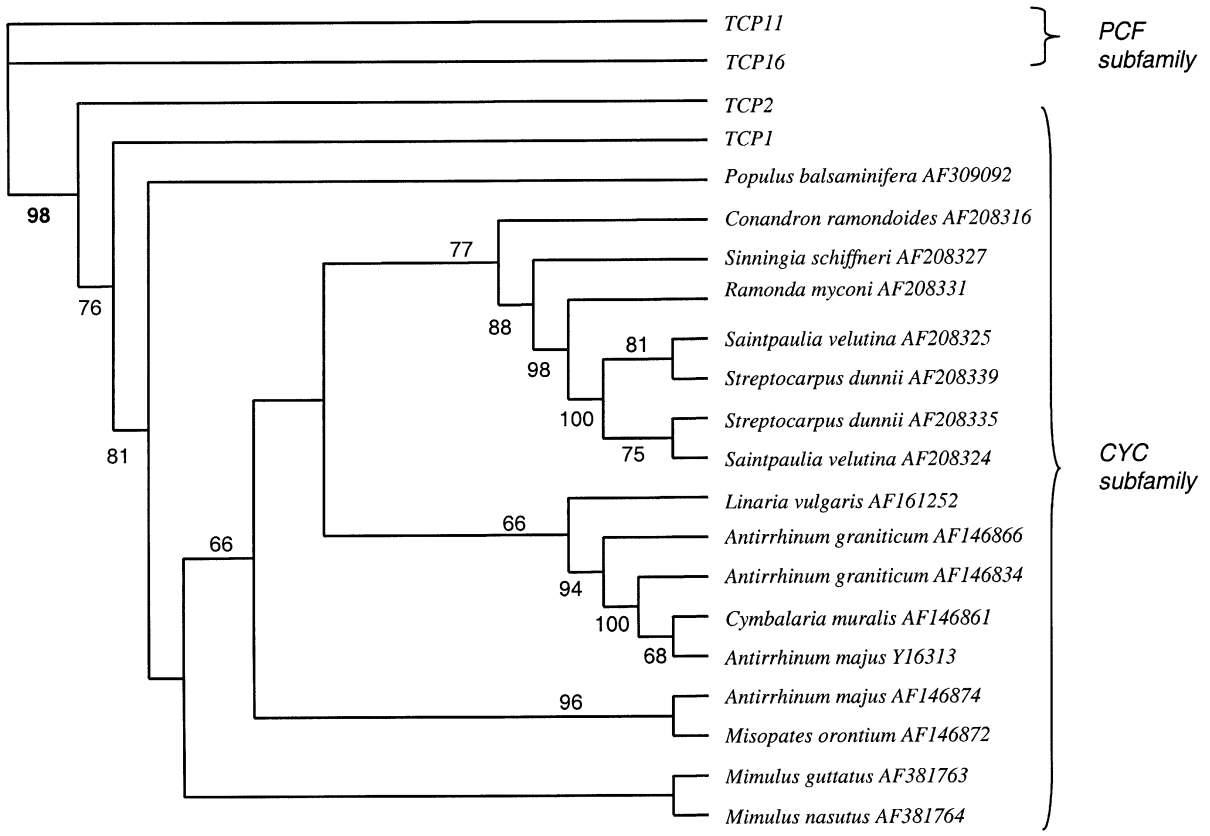

Fig. 2. Consensus maximum parsimony tree obtained from the sample of 21 sequences, using two PCF subfamily sequences as the outgroup. Bootstrap values above 60% are indicated.

Fig. 2. Arbre consensus issu de l'analyse de l'échantillon de 21 séquences par la méthode de parcimonie. Les deux séquences de la sous-famille PCF ont été utilisées comme groupe externe. Les valeurs de *bootstrap* supérieures à 60% sont indiquées.

DICH, *LCYC* and *TCP1*, and a symmetric one for *TCP2* and *TCP3*. The asymmetric expression of *CYC*, *DICH* and *LCYC* is clearly associated with a zygomorphic flower in *Anthirrhinum* and *Linaria* [5, 16, 17]. The transient expression of *TCP1* in *Arabidopsis* may account for both a precocious asymmetry in sepal development [10], and the lack of dorsoventral asymmetry in the fully developed perianth. The primitive function of CYC subfamily genes seems to be related to growth and development of primordia. These genes thus would offer evolution a molecular basis to develop repeatedly asymmetric morphological structures, simply as a consequence of asymmetric expression.

At present, data on expression or function are very scarce from a taxonomic point of view, and sequence data is strongly biased since most data come from Lamiales and *A. thaliana*. Improved knowledge about the evolution of structure and function of CYC subfamily genes will come from a better sampling of various angiosperm taxa.

Acknowledgements

We thank Professors H. Le Guyader and J. Deutsch for critical reading of the manuscript. This work was supported by a grant from the 'Centre national de la recherche scientifique' (France) allocated to C.D.

References

[1] S.F. Altschul, J.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.

[2] Angiosperm phylogeny Group (The), An ordinal classification for the families of flowering plants, Ann. Missouri Botanical Garden 85 (1998) 531–553.

[3] H. Citerne, M. Möller, Q.C.B. Cronk, Diversity of *Cycloidea*-like genes in Gesneriaceae in relation to floral symmetry, Ann. Bot. 86 (2000) 167–176.

[4] P. Cubas, Role of TCP genes in the evolution of morphological characters in Angiosperms, in: Q.C.B. Cronk, R.M. Bateman, J.A. Hawkins (Eds.), Developmental genetics and plant evolution, The Systematics Assoc. Spec. Vol. No. 65 (2002) 247–266.

[5] P. Cubas, C. Vincent, E. Coen, An epigenetic mutation responsible for natural variation in floral symmetry, Nature 401 (1999) 157–161.

[6] P. Cubas, N. Lauter, J. Doebley, E. Coen, The TCP domain: a motif found in proteins regulating plant growth and development, Plant J. 18 (1999) 215–222.

[7] P. Cubas, E. Coen, J.M.M. Zapater, Ancient asymmetries in the evolution of flowers, Curr. Biol 11 (2001) 1050–1052.

[8] J. Doebley, A. Stec, C. Gustus, Teosinte branched1 and the origin of maize: evidence for epistasis and the evoution of dominance, Genetics 141 (1995) 333–346.

[9] J. Doebley, A. Stec, L. Hubbard, The evolution of apical dominance in maize, Nature 386 (1997) 485–488.

[10] P.K. Endress, Evolution and floral diversity: the phylogenetic surroundings of *Arabidopsis* and *Antirrhinum*, Int. J. Plant Sci. 153 (1992) S106–S122.

[11] P.K. Endress, *Antirrhinum* and the Asteridae: evolutionary changes of floral symmetry, Symp. Soc. Exp. Biol. 51 (1997) 133–140.

[12] V. Gaudin, P.A. Lunness, P.R. Fobert, M. Towers, C. Riou-Khamlichi, J.A.H. Murray, E. Coen, J.H. Doonan, The expression of *D-Cyclin* genes defines distinct developmental zones in snapdragon apical meristems and is locally regulated by the *Cycloidea* gene, Plant Physiol. 122 (2000) 1137–1148.

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

[14] T.H. Jukes, C.R. Cantor, Evolution of protein molecules, in: H.N. Munro (Ed.), Mammalian protein metabolism, Academic Press, New York, 1969, pp. 21–132.

[15] L. Lukens, J. Doebley, Molecular evolution of the *Teosinte Branched1* gene among maize and related grasses, Mol. Biol. Evol. 18 (2001) 627–638.

[16] D. Luo, R. Carpenter, C. Vincent, L. Copsey, E. Coen, Origin of floral asymmetry in *Antirrhinum*, Nature 383 (1995) 794–799.

[17] D. Luo, R. Carpenter, L. Copsey, C. Vincent, J. Clark, E. Coen, Control of organ asymmetry in flowers of *Antirrhinum*, Cell 99 (1999) 367–376.

[18] R.G. Olmstead, C.W. DePamphilis, A.D. Wolfe, N.D. Young, W.J. Elisons, P.A. Reeves, Disintegration of the Scrophulariaceae, Am. J. Bot. 88 (2001) 348–361.

[19] D.L. Swofford, PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods), Version 4.06b, Sinauer Associates, Massachusetts, Sunderland, 2001.

[20] C.P Vieira, J. Vieira, D. Charlesworth, Evolution of the Cycloidea gene family in *Antirrhinum* and *Misopates*, Mol. Biol. Evol. 16 (1999) 1474–1483.