Université de Montréal

L'utilité du gène *LEAFY* pour la systématique des Caesalpinioideae (Leguminosae)

par

Annie Archambault

Institut de recherche en biologie végétale
Département de Sciences biologiques
Faculté des Arts et des Sciences

Mémoire présenté à la Faculté des études supérieures en vue de l'obtention du grade de maîtrise en sciences biologiques

Déposé en août 2001

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé :

L'utilité du gène *LEAFY* pour la systématique des Caesalpinioideae (Leguminosae)

présenté par :

Annie Archambault

a été évalué par un jury composé des personnes suivantes :

Bernard Angers	président rapporteur
Anne Bruneau	directrice de recherche
Luc Brouillet	membre du jury

Accepté le 20 novembre 2001

Sommaire

La systématique, qui vise la description et l'explication de la diversité observée dans le monde biologique, s'est dotée de méthodes objectives de reconstruction. Ces méthodes d'analyses phylogénétiques reposent sur des caractères partagés afin de proposer des hypothèses d'évolution des taxons. De plus en plus, les caractères sont issus du génome lui-même. Les systématiciens se sont tournés vers les nucléotides composant la séquence en acide nucléique d'un gène homologue ou les changements structuraux plus rares afin d'accumuler de l'information. Chez les plantes, trois génomes fournissent des caractères moléculaires : le génome chloroplastique, mitochondrial et nucléaire. Le génome chloroplastique est jusqu'ici le plus utilisé alors que le génome nucléaire est sans doute celui dont le potentiel est le moins exploité, cependant plusieurs caractéristiques propres au génome nucléaire en compliquent l'interprétation. Je présente dans un premier chapitre rédigé en français, les particularités de la structure et de l'évolution du génome nucléaire des plantes, illustrées par les performances et difficultés de divers gènes nucléaires, dont fait partie *LEAFY.* La révision concernant le génome nucléaire aidera à mieux comprendre l'intérêt phylogénétique et les risques de l'utilisation de *LEAFY* chez les Caesalpinioideae, l'objectif principal de mon mémoire de maîtrise présenté au second chapitre.

Le génome nucléaire est le plus grand et le plus complexe des génomes des cellules végétales et son évolution est très rapide. L'existence de familles multigéniques et la présence de pseudogènes sont caractéristiques de ce génome, qui ont de plus une incidence sur l'analyse phylogénétique. Les membres des familles multigéniques sont des gènes venant de duplications d'ADN. Les différentes copies n'évoluent pas indépendamment et peuvent être confondues, ce qui rend leur utilisation en analyse phylogénétique ardue. Pourtant, l'intérêt des gènes nucléaires en peu de copies est grand pour la systématique. Une dizaine de ces gènes ont déjà été évalués à cette fin : *RPB2, LEAFY, GAPA* et *GAPB, ADC,* légumines, vicilines, *PHY, 4CL, GBSSI, CYC, ncpGS, PgiC, ADH,* et *CHS* sont du nombre. Ces différents gènes montrent des vitesses d'évolution très variables et la majorité d'entre eux fait partie de petites familles multigéniques où

les copies évoluent indépendamment les unes des autres. Les gènes nucléaires sont aussi caractérisés par la présence d'introns de type épissables. Les théories expliquant l'apparition des introns dans le génome nucléaire sont souvent opposées, mais il semble que l'insertion d'intron soit rare et qu'elle serait un bon marqueur pour caractériser les groupes monophylétiques.

L'objectif du second chapitre, rédigé en anglais sous forme d'article scientifique, est l'évaluation de l'utilité des séquences de *LEAFY* pour la systématique des Caesalpinioideae (Leguminosae). L'histoire évolutive de cette sous-famille n'est pas encore claire, bien qu'ayant été bien étudiée. *LEAFY* est en copie simple dans tous les angiospermes diploïdes étudiées, et il comprend trois introns. Dans les légumineuses étudiées ici, une seule copie de *LEAFY* a été détectée (sauf exception). Cette étude a nécessité le clonage et le séquençage d'une séquence partielle du gène, allant de l'exon 2 à l'exon 3 pour 36 espèces de légumineuses. Des analyses phylogénétiques de parcimonie et de vraisemblance ont été faites avec la portion codante des 74 séquences obtenues. Le signal phylogénétique généré par *LEAFY* est différent de celui de l'intron chloroplastique *trn*L, surtout pour la tribu des Detarieae s. l. Pourtant, dans l'ensemble, les deux marqueurs ont reconnu les mêmes grands groupes monophylétiques. Les hypothèses souvent proposées pour expliquer l'incongruence entre deux marqueurs, soit l'hybridation, le triage de lignées ou la confusion entre paralogues et orthologues, ne semblent pas valides dans ce cas-ci. Les problèmes reliés aux modèles des méthodes d'analyses et à l'échantillonnage restent des causes possibles. Un nouvel intron a été découvert dans des espèces du genre *Brownea*, et pourrait être un bon caractère. La séquence en acides nucléiques du gène nucléaire *LEAFY* procure de l'information phylogénétique à l'échelle taxonomique de la famille et des tribus pour les légumineuses. Plus généralement, cette étude illustre les difficultés de l'utilisation de gènes nucléaires comme *LEAFY* en systématique moléculaire, mais montre aussi que ce génome apporte beaucoup d'informations sur l'évolution des espèces.

Mots clés : *LEAFY*, Caesalpinioideae, génome nucléaire, évolution moléculaire, gène en copie unique, familles multigéniques, analyse phylogénétique, incongruence, introns, caractère.

Table des matières

Liste des figures

Liste des tableaux

Liste des sigles et abréviations

4CL : 4-coumarate : coenzyme A ligase

ADC : Arginine décarboxylase

ADH : Alcool déshydrogénase

ADN : Acide désoxyribonucléique

ARN : Acide ribonucléique

DNA: Deoxyribonucleic acid

GBSSI : Granule-bound starch synthase

bp : bases pair

CHS : Chalcone synthétase

CYC : Cycloidea

GAPC : NAD(P)-glycéraldéhyde 3-phosphate déshydrogénase cytosolique

GAPA et GAPB : deux gènes de la famille multigénique des NAD(P)-glycéraldéhyde 3-phosphate déshydrogénase chloroplastique

kb : kilobases, 1000 paires de bases

lfy: Allèle mutant du gène *LEAFY*

MB : Mégabases (1000 paires de bases)

MP : Maximum parsimony

ML : Maximum likelihood

ncpGS : Glutamine synthétase chloroplastique

PCR : Polymerase chain reaction

pb : paires de bases

pg : picogrammes, 1E-12 gramme

PgiC : Phosphoglucose isomérase

PHY : Phytochromes nommés PHYA, PHYB, PHYC, PHYD, PHYE, etc.

RPB2 : Seconde sous unité de l'ARN polymérase II

RFLP : Random fragment length polymorphism

snRNA : Small nuclear RNA

µl: microlitre, 1E-6 litre

Remerciements

Je remercie tous ceux qui m'ont supporté – dans tous les sens du terme.
Merci d'abord à ma famille, qui m'a inculqué l'émerveillement (papa!), la rigueur (maman!), le travail en équipe (Benoit!) et bien d'autres qualités nécessaires à l'entreprise scientifique. Je vous aime tellement!

Merci mille fois à Anne, tu t'es beaucoup impliquée dans mon projet, mais tu as aussi su me pousser à donner le meilleur de moi, ce dont je ne me croyais même pas capable. Imagine, j'ai fait une maîtrise et même une présentation en anglais! Merci d'avoir eu confiance en moi.

Je remercie mon grand amour, Stéphane, qui m'a fait rire, qui m'a encouragé à me dépasser ou qui m'a consolé, toujours au bon moment. Stéphane, je jure qu'à partir de maintenant, je recommencerai à t'écouter. Merci à tous mes chers amies et amis de tous horizons, qui m'écoutent avec beaucoup de patience lorsque je raconte mes dernières découvertes ou mes nombreuses frustrations. J'adore Christine qui a corrigé tout ce travail avec une gentillesse hors de l'ordinaire.

Je dois beaucoup à M. Michael Frohlich et à M. Jeff Doyle, qui ont accepté ma visite dans leur laboratoire, et dont les discussions scientifiques m'ont énormément stimulée. Je remercie l'Institut de recherche en biologie végétale et le Fond pour la Formation des Chercheurs et l'Aide à la Recherche et ainsi que Anne Bruneau pour leur soutient financier. Je n'aurais pas pu accumuler tant de résultats sans l'aide de mes compagnons de laboratoire : Félix, Josée-Nadia, Fannie, Simon, Marie, Pierre; et ceux des laboratoires voisins aussi – je pense ici à Martin, Sier-Ching, Marie, Frédéric qui m'ont beaucoup aidé en biologie moléculaire.

Chapitre 1

Le génome nucléaire comme source de données sur l'évolution des plantes

Les données moléculaires sont devenues des sources d'informations particulièrement importantes, quasiment incontournables pour étudier la phylogénie des organismes vivants. Les mutations dans les séquences, codantes ou non, provenant des différents génomes, ainsi que les changements structuraux sont des caractères utilisés reconstruire des hypothèses d'évolution. Le présent chapitre présente une revue de littérature des séquences codantes provenant du génome nucléaire utilisées en systématique végétale. Cette section sera précédée d'un survol de la structure et de l'évolution du génome nucléaire chez les végétaux, qui ont une influence sur son utilisation dans les analyses phylogénétiques. Enfin, la découverte de la présence d'un nouvel intron dans quelques espèces d'un genre m'a amené à me pencher sur la fiabilité d'un tel caractère pour élaborer des hypothèses de l'évolution des espèces.

Les génomes des cellules végétales

Chacune des cellules de plantes (sauf les cellules germinales) possède trois génomes non liés : les génomes chloroplastiques, mitochondriaux et nucléaires, et chacun d'eux est soumis à des pressions évolutives différentes. Les deux génomes d'organites cellulaires (chloroplastes et mitochondries) sont présents en multiples copies dans chaque cellule végétale. En effet, dans chacune des cellules, il y a un grand nombre de ces organites, qui abritent chacun des dizaines de copies identiques de génome. Le génome chloroplastique est sans doute le plus connu des trois dans le cas des végétaux. Déjà, la séquence entière du génome chloroplastique est déterminée pour 23 organismes, dont huit angiospermes, une gymnosperme et de nombreuses algues (Organelle Genome Megasequencing Program, 2001). Ce génome fait environ 150 kb chez *Arabidopsis thaliana* (Brassicaceae) et contient 87 gènes (Sato et al., 1999). Les

projets de séquençage de génomes ont révélé qu'un transfert massif de gènes du chloroplaste au noyau s'est produit au cours de l'évolution des plantes (Martin et al., 1998). Le génome chloroplastique possède de nombreuses qualités, en plus d'être relativement court, bien connu et en nombreuses copies, qui en font une excellente source de données phylogénétiques. L'hérédité maternelle et l'absence de recombinaison et de variation allélique facilitent beaucoup l'interprétation de l'évolution du génome chloroplastique tandis que le faible taux de substitution et la présence en multiples copies identiques en facilite techniquement l'étude. Par le passé, les séquences du génome chloroplastique ont donc été largement utilisées en phylogénétique de plantes. Les séquences codantes ont surtout été utiles à de grandes échelles taxonomiques. Par ailleurs, chez les plantes (contrairement aux animaux), les séquences des génomes mitochondriaux subissent infiniment peu de substitutions (Wolfe et al., 1989) alors que la structure est instable et variable (Palmer et al., 2000). Le génome mitochondrial est de ce fait peu utilisé en systématique moléculaire des plantes. Le faible taux de substitution chez ces deux génomes les rend peu propice à l'étude de l'évolution des espèces proches parentes.

Le génome nucléaire est pour sa part très imposant : il a une taille sans commune mesure avec les deux génomes mentionnés plus tôt. Les angiospermes sont le groupe d'organismes vivants dont la taille du génome est la plus variable. D'une espèce à l'autre, la taille peut varier d'un facteur de 1000! La majorité des angiospermes ont cependant un génome moyennement petit, de 0,1 à 3,5 pg d'ADN (100 à 3500 Mb) (Leitch et al., 1998). Lorsque le contenu en ADN est considéré sous l'angle phylogénétique, on détecte des expansions et des contractions de la taille du génome, bien que les angiospermes primitives aient tendance à posséder un petit génome (Leitch et al., 1998). Les duplications multiples et la présence d'éléments mobiles et de séquences répétitives sont des caractéristiques du génome nucléaire qui expliquent partiellement sa grande taille, de même que son évolution rapide, tant en substitutions qu'en réarrangements. Les génomes nucléaires des plantes se distinguent aussi de ceux des vertébrés : ils sont souvent d'origine polyploïde (Masterson, 1994) ou hybride. Les génomes

des plantes ont aussi une plus importante homogénéisation entre les chromosomes, en ce qui a trait au taux de GC par exemple (Schmidt et Heslop-Harrison, 1998). Le génome nucléaire est non seulement imposant, mais sa structure et son histoire sont aussi extrêmement complexes.

Le génome nucléaire : structure et évolution

Il est connu que le génome des eucaryotes supérieurs est très peu compact, particularité qui est exacerbée chez les végétaux. C'est le paradoxe de la valeur C. La valeur C est la constance du contenu en ADN par génome haploïde non-répliqué d'un individu (Bennett et Leitch, 1995). Le paradoxe est que cette valeur n'est pas corrélée à la complexité des organismes, une grande partie de l'ADN semble superflu. Cet ADN « superflu » regroupe les séquences d'introns, de pseudogènes, les éléments mobiles de même que les séquences répétitives centromériques, télomériques ou les séquences hyper variables des minisatellites (Schmidt et Heslop-Harrison, 1998). Les rétrotransposons peuvent expliquer une partie du paradoxe chez les plantes (Federoff, 2000). Ce sont des éléments mobiles : des portions d'ADN capables de se mouvoir d'un endroit à l'autre du génome. Ils sont catégorisés en rétrotransposons (classe I) et en transposons (classe II), et semblent n'avoir pas d'autre fonction que celle de se reproduire et de s'insérer à de plus en plus d'endroits dans le génome. *Arabidopsis,* qui a un très petit génome (130 Mb), abrite peu de copies de rétrotransposons alors que le grand génome du maïs (2 500 Mb) est à moitié composé de ces éléments mobiles (Clegg et al., 1997; Federoff, 2000). Bien que ces séquences semblent n'avoir aucune utilité pour la survie de l'individu, elles jouent un rôle non négligeable dans l'évolution des génomes (Federoff, 2000).

L'une des caractéristiques les plus importantes des angiospermes est qu'ils ont subi plusieurs rondes de polyploïdisation ou de duplications partielles au cours de leur évolution (Masterson, 1994). D'après les séquences du génome complet de la plante modèle *Arabidopsis thaliana*, on sait aujourd'hui que même ce minuscule génome nucléaire aurait subi au moins quatre duplications totales ou partielles (Vision et al., 2000). Puisque ce génome n'est actuellement formé que

de cinq chromosomes, on présume qu'il a aussi subi des fusions de chromosomes et des pertes de gènes massives. Une grande partie des gènes (25% chez *Arabidopsis*) feraient ainsi partie de familles multigéniques (Vision et al., 2000). La divergence entre les copies de même que le nombre de copies est extrêmement variable. Par exemple, certains gènes comme l'*ADH* font partie de petites familles multigéniques de deux à sept membres (Clegg et al., 1997) alors que d'autres familles comme celle des gènes MADS-box ont jusqu'à 50 membres différents (Alvarez-Buylla et al., 2000). Le cas des ADN ribosomaux, qui existent en milliers de copies, est particulier puisque les copies sont presque toutes semblables : elle ont subi de l'évolution concertée (Baldwin et al., 1995). Les gènes dupliqués peuvent être répétés en tandem (l'un à la suite de l'autre) ou se trouver non-liés, sur des chromosomes différents.

L'évolution des régions dupliquées

La différence entre gènes en copie simple ou en copies multiples semble triviale, mais il s'agit d'une distinction cruciale pour l'évolution des gènes. Les gènes en copies multiples peuvent être soumis à des mécanismes d'évolution différents de ceux impliqués pour les gènes en copie simple.

1) Les gènes dupliqués pourront poursuivre leur évolution de façon autonome et continuer d'être transcrits et fonctionnels. Par accumulation progressive de mutations neutres et non-délétères, ils divergeront l'un de l'autre peu à peu. Le degré de divergence entre deux copies est très variable et dépend de plusieurs facteurs dont l'âge de la duplication et l'importance de la protéine en question pour l'organisme.

a) Ils peuvent demeurer presque identiques et accomplir une fonction similaire, on dit que les gènes sont redondants. Les gènes *SEPALLATA 1 2* et *3* constituent un bel exemple de redondance (Pelaz et al., 2000). La séquence de ces trois gènes MADS-box est semblable, mais tout de même différenciée. La perte de fonction de l'un des gènes *SEPALLATA* n'affecte pas le développement. Par contre, lorsque les trois gènes sont inactivés, les organes floraux produit ne sont que des sépales (Pelaz et al., 2000).

b) Les gènes dupliqués peuvent diverger énormément, jusqu'à ce qu'il soit même difficile de percevoir leur origine commune. Il a couramment été observé que dans les lignées suivant une duplication, le taux de substitution non-synonyme s'accélère. Cette situation est propice à l'acquisition de nouvelles fonctions par l'une des copies. C'est le cas des stilbène synthétases, qui ont été recrutées indépendamment plusieurs fois, toujours à la suite de duplications du gène chalcone synthétase (Helariutta et al., 1996; Durbin et al., 2000). Ce mécanisme de duplication et de divergence a été proposé comme la force majeure d'évolution des génomes des eucaryotes (Ohno, 1970). Lorsque les gènes ou génomes sont dupliqués, l'une des copies pourrait « expérimenter » de nouvelles fonctions alors que l'autre continuerait d'exercer sa fonction initiale.

c) La duplication pourra être suivie de la mise en silence de l'une des copies dupliquées (*gene silencing*). Plusieurs mécanismes sont actuellement proposés pour expliquer la mise en silence des gènes. Certains peuvent impliquer la méthylation de l'ADN, d'autres, l'interférence entre les ARN double brins ou l'altération de la chromatine (Hsieh et Fire, 2000). Plus couramment, cette forme d'évolution se produit lorsque l'une des copies accumule des mutations qui rendent la protéine produite dysfonctionnelle, des mutation stop ou qui affectent le cadre de lecture par exemple. Ces copies sont alors devenues des pseudogènes. Dans l'évolution des vertébrés, 50% duplications mèneraient à la formation de pseudogènes (Wagner, 1998).

2) Les gènes dupliqués peuvent évoluer de concert, de façon beaucoup plus intime en subissant de la recombinaison.

a) Il peut y avoir des crossing-over entre les copies, par un échange de segment entre allèles ou entre loci. Lorsque l'échange de segments est entre loci, les gènes ainsi formés sont composites et leur histoire est réticulée puisque les différentes régions du gène ne partagent pas la même histoire. Les crossing-over ont habituellement lieu sur une plus longue région que la conversion génique (Berry et Barbadilla, 2000).

b) Ou encore, une copie peut en convertir une autre en sa séquence ou en une partie de sa séquence. La conversion génique est l'un des mécanismes d'évolution moléculaire ayant été proposés pour expliquer l'homogénéisation des

copies des gènes d'ARN ribosomaux (Baldwin et al., 1995), mais dont l'impact reste sous-évalué pour l'évolution des autres gènes. Les méthodes de comparaison de séquences ont révélé que la conversion de gènes a eu lieu dans l'histoire des familles multigéniques des chalcone synthétases de *Ipomea* (Huttley et al., 1997) et dans les gènes d'actines des angiospermes (Moniz de Sa et Drouin, 1996), entre autres exemples.

La dynamique et les processus d'évolution qui règlent la taille des familles multigéniques sont encore mal compris. Le nombre de copies de plusieurs familles multigéniques pourrait être en constante fluctuation où l'évolution serait caractérisée par de récentes duplications en parallèle (Clegg et al., 1997; Small et Wendel, 2000). On s'explique mal comment certaines familles peuvent être si imposantes alors que d'autres semblent être limitées à moins de cinq copies (Clegg et al., 1997). De plus, les membres d'une famille multigénique peuvent interagir entre eux toutes ces façons (mise en silence, divergence, recombinaison, conversion génique) sans que l'on sache véritablement les raisons favorisant tels phénomènes par rapport aux autres. Même s'il est reconnu que les éléments mobiles (Federoff, 2000) ou les duplications stimulent l'évolution des génomes, il n'est cependant pas encore très clair comment tous ces facteurs affectent le phénotype des organismes.

Alors que l'on reconnaît la grande vitesse d'évolution des génomes et l'effet des éléments mobiles, c'est la stabilité des génomes et la synténie qui étonne davantage. En effet, l'ordre des gènes et même la position des introns sont étonnamment conservés entre espèces (la synténie) souvent taxonomiquement éloignées (Federoff, 2000), bien que les régions intergéniques varient grandement en taille et en séquence. La synténie procure certains avantages au systématicien moléculaire : l'amplification de grandes régions nucléaires par PCR est alors facilitée.

Les substitutions : les caractères de l'évolution moléculaire

Traditionnellement, les substitutions, ou mutations ponctuelles, sont les mutations les plus couramment utilisées pour étudier l'évolution des gènes orthologues dans différentes espèces ou des paralogues d'une famille multigénique. Beaucoup d'autres mutations se produisent couramment, mais ne sont pas encore autant utilisées dans des études phylogénétiques.

Le génome nucléaire est un choix évident pour trouver un grand nombre de caractères variables. En effet, le dynamisme quasi frénétique des génomes nucléaires des végétaux génère un taux de substitution extraordinairement élevé comparativement aux génomes chloroplastiques ou mitochondriaux (Wolfe et al., 1989). La redondance partielle et la recombinaison entre allèles et entre locus sont souvent invoquées pour expliquer cette si haute variabilité. Les génomes nucléaires montrent aussi une grande variabilité du taux d'évolution : certaines régions du génome subissent beaucoup de substitutions alors que d'autres changent peu (Wolfe et al., 1989).

Les différents sites d'une séquence codante évoluent aussi à des taux de substitution très variés. À cause de la dégénérescence du code génétique, certaines positions du codon, la troisième position notamment, peuvent varier sans changer l'acide aminé codé. Ce sont des substitutions synonymes. Dans un gène codant, on prend habituellement pour acquis que les substitutions synonymes ne sont soumises à aucune sélection, on utilise le taux de substitution synonyme comme étant représentatif ou égal au taux de mutation neutre, sans l'action de la sélection (Li, 1997). Les différents sites de la séquence d'un gène n'ont donc pas tous le même taux de substitution, certains ne sont soumis à aucune sélection purifiante alors que pour d'autres sites, toute substitution est létale. Les taux de substitution non-synomyme sont extrêmement variables d'un gène à l'autre (Li, 1997) dans un même génome alors que les taux de substitution synonyme varient aussi (Wolfe et al., 1989), mais d'une façon beaucoup moins dramatique. Ces différences de taux de substitution sont souvent pris en

considération dans les modèles d'évolution des méthodes analyses phylogénétiques.

De nouvelles possibilités pour les séquences nucléaires

Malgré la complexité de l'évolution du génome nucléaire, les chercheurs se tournent de plus en plus vers le génome nucléaire comme source de données pour retracer l'évolution des espèces. Cet intérêt est motivé par l'idée générale que la diversité des phénotypes trouve ses racines à même le génome nucléaire. De cette même idée sont aussi nés les nombreux projets de séquençage de génomes (National Center for Biotechnology Information, 2001a). En outre, la base génétique de nombreux processus biologiques, comme la régulation du développement, la réponse aux stress, la résistance aux pathogènes la communication inter-cellulaire, est de mieux en mieux comprise.

Plusieurs études s'intéressent maintenant à la comparaison des différents modes de développement dans des espèces apparentées en comparant la séquence et l'expression des gènes impliqués dans le développement des différents organes (Hofer et Ellis, 1998; Citerne et al., 2000), qui surtout des facteurs de transcription ou des facteurs de transduction. C'est dans cette perspective que de plus en plus de systématiciens moléculaires se tournent vers les séquences du génome nucléaire pour retracer l'évolution des espèces végétales qui les intéressent.

L'arbre des gènes : reflet de l'arbre des espèces ?

D'un point de vue théorique, un génome, quel qu'il soit, doit posséder des caractéristiques particulières afin de fournir des données phylogénétiques utiles à l'étude de l'évolution des espèces. Il ne faut jamais perdre de vue que la phylogénie d'un gène présent dans différentes espèces n'est pas nécessairement équivalente à celle de la phylogénie des espèces. Une analyse phylogénétique d'un gène nucléaire doit faire l'objet d'attentions particulières à cause des mécanismes d'évolution génétique qui sont propres au génome nucléaire et que

nous avons déjà survolés. Nos connaissances limitées de ce génome font en sorte qu'il est difficile de déterminer quels sont les gènes et régions qui pourraient procurer un contenu phylogénétique fiable.

Penchons-nous les précautions nécessaires pour que la séquence d'un gène procure un contenu phylogénétique fiable, qui puisse représenter l'histoire de l'organisme qui le possède.

1) L'évolution du gène doit être telle qu'elle puisse être analysée par les méthodes et logiciels de reconstruction d'arbres phylogénétiques. En effet, des méthodes d'analyses sont incapables d'analyser correctement des séquences dans certaines situations bien précises (Felsenstein, 1978; Swofford et al., 1996; Siddal, 1998). L'échantillonnage en taxons, l'histoire du gène et la méthode d'analyse doivent tous être pris en considération lors d'une analyse phylogénétique, puisque tous ces facteurs sont reliés.

L'échantillonnage doit être fait en fonction de la variabilité du gène. Obtenir le niveau adéquat de variation est crucial. En effet, lorsque des séquences ont divergé depuis très longtemps ou à un rythme accéléré, elles ont une probabilité élevée d'avoir subi des substitutions multiples à un site. Cette situation se produit lorsqu'un site a connu tellement de mutations dans son histoire, que le même état (par exemple un A) à ce site dans deux séquences différentes pourrait fort bien ne pas être homologue (Swofford et al., 1996). Le site aurait pu devenir un T, puis subir une nouvelle mutation vers un A, on conclut alors qu'il y a saturation de substitutions. Les substitutions multiples sont, on le comprendra, néfastes à l'analyse phylogénétique. Pour savoir à quel niveaux taxonomiques le gène peut procurer de l'information, on doit connaître le taux relatif de variation entre taxons. Malheureusement, des lignes directrices précises n'existent pas pour établir le lien entre le taux de substitution et le niveau d'utilité. Il faut donc, soit connaître d'avance d'après la littérature, soit établir par des analyses préliminaires, à quel niveau taxonomique il y a suffisamment de variabilité pour procurer de la résolution, mais pas trop pour brouiller l'analyse.

Le modèle d'évolution intrinsèque à la méthode d'analyse choisie doit aussi être compatible avec le patron de substitutions du gène nucléaire. Le modèle d'évolution doit permettre que le taux d'évolution entre lignées varie puisque les gènes nucléaires subissent souvent une accélération de leur évolution à la suite, par exemple, d'une duplication (Durbin et al., 2000), ou d'un temps de génération accéléré dans une lignée (Gaut et al., 1996). Un problème courant découlant de cette situation est celui de l'attraction des longues branches dans les analyses cladistiques. Ce problème se produit lorsque deux lignées non reliées évolutivement subissent une évolution plus rapide que les autres lignées. Dans des analyses cladistiques, ces deux lignées peuvent se regrouper faussement (Felsenstein, 1978; Wendel et Doyle, 1998). Il est aussi reconnu que les différentes positions du codon n'évoluent pas au même rythme puisqu'elles ne sont pas soumises à la même pression de sélection. Ces variations du taux de substitution entre sites doivent pouvoir être analysées par les modèles de la méthode choisie (Yang, 1996). L'hétérogénéité du taux de substitution entre sites ou entre lignées, du taux de GC ou du taux de transition/transversion sont toutes des situations qui peuvent être critiques pour l'analyses. Ce sont toutes des situations difficiles à traiter par les logiciels d'analyse, et elles devraient donc être évitées autant que possible en choisissant soigneusement la qualité et la quantité de l'échantillonnage (Graybeal, 1998).

2) La fréquence de l'hybridation et de l'introgression chez les plantes (Xu, 2000) sont des facteurs qui obligent à prendre des précautions particulières lors d'analyses incluant des gènes nucléaires. On voudra connaître la séquence de chacun des allèles d'un gène, surtout dans les études qui incluent l'hybride et ses espèces parentales. L'hybridation peut surtout être un problème à de petites échelles taxonomiques.

3) On doit avoir l'assurance de comparer deux gènes homologues dans les espèces étudiées. Le gène doit donc être idéalement en copie unique. Si le gène fait partie d'une famille multigénique, on doit avoir une connaissance assez poussée de la famille afin de pouvoir distinguer aisément les différentes copies du

gène. Une terminologie particulière a été élaborée pour décrire les relations d'homologie entre les membres d'une famille multigénique par Fitch (voir Doyle et Davis, 1998). Les copies d'un gène issues de duplications sont appelées des paralogues alors que la spéciation mène à l'existence de gènes orthologues, présents dans des espèces différentes. Ainsi, deux gènes très similaires retrouvés à l'intérieur d'un même individu ne peuvent être des orthologues, mais pourraient être des paralogues. Une confusion entre paralogues et orthologues peut conduire à des erreurs d'interprétation des relations entre espèces dans une analyse phylogénétique, surtout si l'échantillonnage est incomplet. Supposons que dans une espèce ancestrale a eu lieu une duplication menant aux gènes A et B dans le même génome. Cette espèce ancestrale a par la suite donné naissance à trois espèces : 3, puis 2 et 1 (figure 1.1 A). Les gènes A et B peuvent indépendamment représenter les spéciations. Par contre, si la duplication n'est pas été détectée et qu'une analyse comporte seulement les gènes 1A, 2B et 3A (figure 1.1 B), on commettra l'erreur de considérer les espèces 1 et 3 comme groupe monophylétique et l'espèce 2 comme groupe-frère alors que ce n'est pas le cas. La terminologie n'est à ce jour pas assez

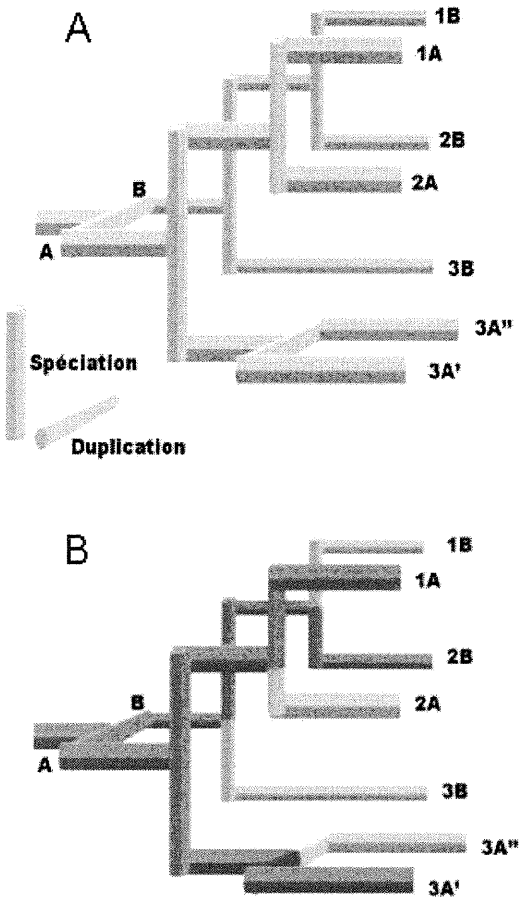


Figure 1.1 Schéma représentant les relations d'orthologie et de paralogie entre des séquences lorsque ont eu lieu deux duplications, menant aux gènes A et B, puis A' et A"; et deux spéciations, menant aux espèces 1, 2 et 3. (A) Un échantillonnage complet de paralogues et d'orthologues. (B) Un échantillonnage partiel (en foncé) de gènes ne comprenant que 1A, 2B et 3A.

raffinée pour distinguer, par exemple, la relation entre 2A et 3A' ou 2A et 3A" (Doyle et Davis, 1998).

Afin d'éviter de telles erreurs d'interprétation lors d'analyses phylogénétiques avec des membres de familles multigéniques, il faut avoir l'assurance de comparer des orthologues. Pour ce faire, l'échantillonnage devrait être aussi complet que possible, composé de tous les paralogues et orthologues. Ce n'est malheureusement pas toujours possible, certains membres ayant pu être effacés du génome d'une espèce (la perte de gène). L'échantillonnage exhaustif n'est toutefois pas suffisant pour établir une solide preuve d'orthologie. La comparaison du lieu d'expression et de la fonction d'un gène peut faciliter cette tâche. Des exemples provenant de la famille de gènes MADS-box ont montré que les gènes orthologues ont tendance à posséder le même patron d'expression et à assumer la même fonction dans des espèces végétales différentes (Hasebe et Banks, 1997; Purugganan, 1998). La preuve la plus convaincante d'orthologie réside en la connaissance de la carte génétique des génomes étudiés. C'est seulement lorsque l'on peut prouver que les régions chromosomiques – gènes ou régions intergéniques - entourant le gène à l'étude sont les même d'une espèce à l'autre que l'on prouve véritablement le lien d'orthologie entre des gènes (Small et Wendel, 2000).

4) Même lorsque la confusion entre paralogues et orthologues ne pose pas de problèmes, d'autres complications peuvent survenir. Idéalement, l'évolution des gènes à l'étude devrait être faite uniquement de divergences en embranchements dichotomiques, mais lorsque le gène étudié est en multiples copies ou qu'il fait partie d'une famille multigénique, les différentes copies peuvent s'être échangé des segments d'ADN. Ces régions du gène n'ont alors pas la même histoire, on dira que leur évolution est réticulée. À l'heure actuelle, plusieurs logiciels spécialisés peuvent détecter l'évolution réticulée (McDade, 1995; Rieseberg et Morefield, 1995), mais les logiciels les plus puissants, pouvant analyser un grand nombre de séquences selon diverses modèles d'évolution, ne le peuvent toujours pas. Les processus de réticulation ne sont pas aisément

détectables, mais peuvent être reconnus par la comparaison attentive des séquences dans l'alignement, par leur comportement dans l'analyse phylogénétique ou par d'autres méthodes encore (Drouin et al., 1999). En général, lorsque la longueur, la position et la séquence d'introns entre les paralogues sont très différentes, on peut considérer la conversion génique peu probable. La conversion génique a été détectée dans de nombreuses familles multigéniques, par exemple, entre les paralogues des actines (Moniz de Sa et Drouin, 1996; Drouin et al., 1999), des chalcone synthétases (Huttley et al., 1997) et des alcools déshydrogénase dans certaines espèces (Gaut et al., 1999); le gène *PgiC* est aussi souvent d'origine recombinante (Liu et al., 1999). Les gènes d'origine recombinante ne devraient pas être utilisés tels quels pour des analyses phylogénétiques. Si c'était inévitable, les régions d'origines différentes devraient être analysées séparément. Voilà pourquoi l'idéal demeure de travailler avec des gènes en copie strictement unique.

Il est toujours préférable d'évaluer l'utilité d'un marqueur phylogénétique au préalable, surtout lorsqu'il s'agit d'un gène nucléaire. Cette évaluation se fait traditionnellement par la comparaison de la topologie obtenue par ce nouveau gène avec la topologie acceptée par les spécialistes du groupe à l'étude ou celle d'autres données dont la fiabilité est éprouvée. Concrètement, cette comparaison se fait souvent avec des gènes ou régions non-codantes chloroplastiques puisqu'elles sont très courantes. La concordance entre des données indépendantes est un important critère de la robustesse de l'hypothèse d'évolution des espèces (Friedlander et al., 1996). Lorsque de l'incongruence entre des ensembles de données se produit, les hypothèses d'évolution des espèces sont plus difficiles à tester, mais c'est souvent là que l'évolution moléculaire devient la plus intéressante!

Gènes nucléaires utilisés en systématique

La suite de cette révision consiste en un survol des gènes nucléaires (autres que les gènes ribosomaux) actuellement facilement disponibles pour construire des phylogénies d'espèces de plantes. Ils ne représentent qu'une infime fraction de toutes les séquences codantes du génome nucléaire, mais en montrent toute la complexité. Parce qu'ils ont en milliers de copies et que leur structure est conservée, les gènes ribosomaux ont été beaucoup utilisés dans les débuts de la systématique moléculaire (pour une révision, voir par exemple Baldwin et al. (1995)). Afin d'obtenir des informations supplémentaires du génome nucléaire et parce que les séquences ribosomales ne peuvent pas être utiles à tous les niveaux taxonomiques, les systématiciens ont cherché à confirmer leurs hypothèses avec d'autres loci. À ce jour, l'on compte environ une dizaine de gènes en copie unique ou en peu de copies dont l'utilité a été validée en systématique végétale. L'objectif de cette section est de faire une mise à jour de l'application des gènes nucléaires aux fins d'étude phylogénétiques. Les principales difficultés rencontrées lors de l'utilisation de ces séquences seront mentionnées. Les gènes seront présentés en ordre décroissant approximatif d'échelle taxonomique.

ARN polymérase II

L'ARN polymérase II catalyse la synthèse d'ARN messager chez les eucaryotes. La sous-unité RPB2 est la seconde plus grande de tout le complexe protéique. Elle pèse 140 kDa, et le gène est fait de 3,5kb de séquences codantes et est interrompu par 24 introns (Figure 1.2) (Denton et al., 1998). Comme la protéine doit être en contact avec plusieurs autres unités dans le complexe protéique, sa structure est très conservée (Denton et al., 1998). L'analyse des séquences de *RPB2* a d'ailleurs déjà contribué à résoudre des questions d'évolution des lignées d'eucaryotes (Sidow et Thomas, 1994). Dans une étude comprenant une dizaine d'espèces, Denton et al. (1998) ont montré que l'architecture du gène est conservée à travers les plantes vertes. Les séquences partielles obtenues montrent que la position et le nombre d'introns sont constants,

quoique leur séquence varie grandement. Les premières études montraient, par hybridation Southern, par RFLP, par test mendélien et par séquençage de génome, que le gène *RPB2* est en copie unique chez les angiospermes étudiés – *Arabidopsis, Solanum, Rhododendron* (Denton et al., 1998). Par contre, une étude ultérieure a mis en évidence deux copies de *RPB2* dans les Gentianales (Oxelman et Bremer, 2000). L'une des copies ne possédait aucun intron dans la séquence partielle étudiée. Alors que l'analyse fondée sur un très petit échantillonnage retrouvait facilement les grands clades des plantes vertes (Denton et al., 1998), une analyse au niveau familial et ordinal, comprenant plus de 30 espèces de Gentianales n'était pas congruente avec les relations telles qu'inférées par le génome chloroplastique (Oxelman et Bremer, 2000).

FLORICAULA / LEAFY

La protéine LEAFY est un facteur de transcription impliqué dans le développement floral. La présence de la protéine est nécessaire, mais pas suffisante, pour activer la transcription de plusieurs gènes permettant au méristème inflorescenciel de produire des fleurs (Parcy et al., 1998). Dans le cadre d'une étude comprenant environ 20 espèces, Frohlich et Parker (2000) ont montré que la séquence en acides aminés de LEAFY est assez conservée pour qu'il soit possible de comparer des séquences de fougères avec celles de gymnospermes et d'angiospermes évoluées, qui possèdent toutes trois exons (Figure 1.2). Néanmoins, le gène, dont la séquence codante mesure un peu mois de 1,5 kb, possède des régions de haute variabilité. Il a aussi été démontré qu'une seule copie de *LEAFY* est transcrite pour tous les angiospermes diploïdes, ce qui est très rares chez les gènes nucléaires (Frohlich et Meyerowitz, 1997; Frohlich et Parker, 2000). Par contre, des pseudogènes ont été découverts chez certaines espèces (Frohlich et Parker, 2000). Cette phylogénie de séquences en acides aminés, comprenant plus de 20 espèces, a pu mettre en évidence les relations maintenant acceptées entre les gymnospermes et les angiospermes, et ce avec de hautes valeurs de support (Frohlich et Parker, 2000).

Glycéraldéhyde-3-phosphate déshydrogénase

NAD(P)-GAPDH (NAD(P)-glycéraldéhyde 3-phosphate déshydrogénase) est une enzyme glycolique du cytosol codée par le gène *GAPC*. Ce gène fait partie d'une petite famille multigénique avec les gènes *GAPA* et *GAPB* (Pohlmeyer et al., 1996). Ces deux derniers sont codés dans le génome nucléaire, mais l'enzyme est ensuite transportée au chloroplaste, et ils sont plus proches parents l'un de l'autre que du gène codant pour l'enzyme cytosolique *GAPC*. *Arabidopsis* ne possède qu'une seule copie de *GAPC* alors que le maïs en a trois (Martin et al., 1993). Une analyse avec près de 1000 pb de séquences codantes de *GAPC* a été effectuée pour une quinzaine d'espèces de plantes vertes parmi les bryophytes, les gymnospermes et les angiospermes (Martin et al., 1993). Le taux de substitution de *GAPC* est faible, l'évolution du gène semble lente et les séquences sont comparables entre règnes jusqu'à un certain point. La topologie obtenue par Martin et al. (1993) n'est cependant pas en accord avec celle couramment acceptée aujourd'hui. Le gène *GAPC* pourrait être d'une certaine utilité à de grands niveaux taxonomiques, mais un échantillonnage plus étoffé sera nécessaire pour le montrer.

Les unités *GAPA* et *GAPB* sont codées par des gènes nucléaires plutôt courts formant des unités de 36 et 39 kDa respectivement (Pohlmeyer et al., 1996). Une seule analyse phylogénétique des séquences du gène *GAPA* a été réalisée. D'après des séquences partielles d'environ 660 pb de *GAPA* chez seulement neuf espèces de la famille des Berberidaceae, il semble qu'il y ait congruence entre les phylogénies de *GAPA* et de *rbc*L (Adachi et al., 1995). L'échantillonnage réduit diminue toutefois la crédibilité de l'étude.

Arginine décarboxylase

L'arginine décarboxylase (*ADC*) est l'une des enzymes de la voie biochimique de la production des putrescines et polyamines chez les plantes; elle doit former un dimère pour être active. L'activité de l'ADC est corrélée aux stress environnementaux, à la floraison et au développement floral (Galloway et al.,

1998). Ce gène n'est pas exclusif aux plantes, on le retrouve chez les bactéries, les animaux et possiblement les mycètes et les protistes aussi, mais pas chez la levure (*S. cerevisiae)*. L'*ADC* est codé par 2100 pb non interrompues par des introns (Figure 1.2). Il existe deux loci de L'*ADC* chez *Arabidopsis* (Watson et al., 1997), *ADC1* et *ADC2*. Une étude évaluant l'utilité phylogénétique de l'*ADC* dans les Brassicaceae (Galloway et al., 1998) a démontré que le gène est moyennement variable : le taux de substitution non-synonyme pour onze genres est comparable à celui de *ndh*F du génome chloroplastique alors que le taux de substitution synonyme en est cinq fois plus élevé. Dans cette même étude, les relations proposées par l'*ADC* étaient dans l'ensemble congruentes avec *ndh*F et *rbc*L, et elles étaient bien supportées. Il en a été conclu que le gène *ADC* pourrait être un excellent marqueur moléculaire pour établir les relations à l'intérieur et entre familles de plantes.

Legumines et vicilines

Il existe deux types de protéines d'entreposage faisant partie d'une classe de protéines appelée globulines : les légumines et les vicilines. Les globulines de type 11S sont appelées légumines alors que l'on appelle les globulines 7S vicilines. Ces protéines sont tenues pour responsables de beaucoup de réactions allergiques aux arachides chez les humains (Maleki et al., 2000). Dans plusieurs groupes, dont les légumineuses, les gènes des légumines sont en copies multiples, nommées *LEGA, LEGB, LEGC* (notées *legA, legB, legC* dans l'article original*)* etc. La séquence des légumines comprend un signal peptidique, puis environ 1500 pb de séquences codantes. Un rapide survol dans les bases de données de séquences (National Center for Biotechnology Information, 2001b) révèle que les légumines sont souvent interrompues par trois introns. L'utilité taxonomique d'une séquence partielle des légumines a été évaluée. Avec un échantillonnage de moins de dix espèces, une analyse a pu retrouver le clade des angiospermes (Lang et Fisher, 1995). Toutefois, peut-être à cause du faible échantillonnage, les relations à l'intérieur des angiospermes sont parfois douteuses, quoique bien supportées. D'autre part, les données sérologiques des légumines étaient assez informatives pour procurer des regroupements fiables,

dans la famille des Ranunculaceae (Jensen, 1995). Cette dernière étude nous permet d'espérer que les séquences en acide nucléique des légumines sont assez variables pour être utiles en systématique a l'échelle taxonomique des familles.

Les gènes vicilines (les globulines 7S) forment une petite famille multigénique codant pour des protéines d'entreposage chez les plantes à fleurs. Ces gènes comprennent environ 2200 pb, habituellement entrecoupées de cinq introns d'une centaine de paires de bases chacun (Figure 1.2). Une étude portant sur les relations génériques au sein de la famille des Sterculiaceae (Whitlock et Baum, 1999) a montré, en utilisant environ 720 pb d'exon, que le gène viciline est présent en copie unique dans cette famille et accumule les changement environ 5 à 10 fois plus vite que le gène *ndh*F. Ce rythme est approprié pour l'analyse des relations au niveau intergénérique. L'arbre des espèces suggéré par ce gène était plausible (Whitlock et Baum, 1999). Cependant, le génome des légumineuses contient plusieurs copies de vicilines, souvent plus de cinq. Dans le genre *Lens* des légumineuses, au moins quatre copies vicilines ont été reportées (Saenz de Miera et Pérez de la Vega, 1998). Ces différentes copies ne montraient pas de signe d'évolution concertée. Il est donc probable que l'on pourra reconnaître les orthologues des paralogues dans une analyse phylogénétique, mais il reste que l'inférence à la phylogénie des espèces en serait plus risquée.

Phytochromes

Les phytochromes sont des photorécepteurs de la lumière rouge et infrarouge présents dans toutes les plantes vertes (Mathews et al., 1995). Ces grands récepteurs protéiques, formés de plus de 1000 acides aminés, sont attachés de façon covalente à un chromophore tetrapyrrole. La perception des changements de condition lumineuse par les photorécepteurs a un effet sur diverses réponses du développement cruciales pour le cycle de vie, comme la germination, le développement des organites, la synthèse des flavonoïdes et l'induction florale (Mathews et al., 1995).

Les gènes codants pour les phytochromes font partie d'une famille multigénique plus ou moins élargie selon le groupe de plantes. Il n'y a en fait qu'une ou deux copies dans la plupart des ptéridophytes et gymnospermes alors que cinq gènes de phytochromes sont présents dans les espèces diploïdes d'angiospermes. *Arabidopsis thaliana* possède cinq phytochromes nommés *PHYA* à *PHYE*, localisés sur quatre chromosomes différents, ils sont tous constitués d'un grand exon de plus de 2000 pb suivi de trois plus petits exons (Figure 1.2) (Mathews et al., 1995). Il y a certainement eu des duplications et des pertes de gènes dans les lignées récentes parce que différentes espèces d'angiospermes ne possèdent pas toutes chacune des copies (Mathews et al., 1995) et d'autres en ont plus de cinq. D'après les analyses phylogénétiques de la famille multigénique des phytochromes dans les angiospermes et les extra-groupes, il semble que deux duplications majeures aient eut lieu. Une première duplication se serait produite après la séparation des fougères et des autres plantes terrestres menant aux classes *PHYA/C* et *PHYB/E* (Donoghue et Mathews, 1998). La seconde duplication majeure aurait eu lieu après l'apparition des angiospermes, mais avant la séparation des monocotylédones et des eudicotylédones, ayant mené aux classes *PHYA* et *PHYC* (Donoghue et Mathews, 1998). Plusieurs duplications ont aussi eu lieu dans des groupes restreints, à l'intérieur de familles ou d'ordres particuliers.

Les phytochromes ont été utiles dans de nombreuses études de systématique végétale. C'est par l'analyse phylogénétique des séquences de 1 à 2 kb de l'exon 1 des gènes *PHYA* et *PHYC* dans 26 espèces d'angiospermes primitives que Mathews et Donoghue (1999) ont pu identifier la première lignée des angiospermes, celle menant à *Amborella*. Les différentes classes des *PHY* ont aussi été utilisées pour résoudre l'évolution à l'intérieur de familles telles que les Leguminosae (Papilionoideae), les Poaceae et les Celastraceae. Dans les deux derniers cas, 1 kb de l'exon 1 du phytochrome B a été séquencé (Mathews et al., 2000; Simmons et al., 2001) et un peu moins pour l'étude sur les légumineuses (Lavin et al., 1998). Dans l'étude de 51 espèces de Poaceae, les séquences du phytochromes B ont confirmé, avec de bonnes valeurs de support,

de précédentes hypothèses fondées sur la morphologie et les séquences chloroplastiques, et les sous-familles principales ont été reconnues (Mathews et al., 2000). Pour les Celastraceae, l'analyse avec 51 espèces avait aussi démontré qu'il n'y avait pas de duplication de *PHYB* à l'intérieur de cette famille (Simmons et al., 2001). Dans cette analyse, *PHYB* s'est montré très efficace pour définir des groupes proches parents alors que les branches plus lointaines étaient instables et peu supportées. *PHYB* a tout de même confirmé que les tribus et groupes des Celastraceae, tels que définis dans les années 1940, devaient être abandonnés. Dans ces deux études, l'appartenance à une famille multigénique du gène à l'étude n'a pas crée de problèmes de paralogie. Dans l'étude sur les légumineuses, 113 séquences de 3 classes de phytochrome (*PHYB*, *PHYE*, *PHYA1* et *PHYA)* chez cinquante espèces ont été étudiées (Lavin et al., 1998). La copie orthologue à *PHYB* a aussi été retrouvée, mais chez très peu d'espèces, et aucune copie de *PHYC* n'a été amplifiée (Lavin et al., 1998). Les relations entre les espèces proposées par les copies de *PHYE*, de *PHYA* et de *PHYA1* considérés séparément étaient généralement congruentes et crédibles.

Il semble généralement que les paralogues de la famille multigénique des phytochromes puissent être facilement différenciés les uns des autres et qu'ils possèdent un taux de variation approprié pour l'étude des relations entre genres à l'intérieur d'une famille. L'évolution des différentes copies ne se fait pas au même rythme, pas plus qu'elle ne suit une horloge moléculaire entre les lignées (Mathews et al., 1995), mais il semble que ce *tempo* d'évolution n'affecte pas la possibilité d'utiliser les phytochromes comme marqueurs phylogénétiques.

4-coumarate : coenzyme A ligase

Dans la voie de biosynthèse de la lignine, l'enzyme codée dans le noyau, la 4-coumarate : coenzyme A ligase (*4CL*) fait partie d'une famille multigénique et a déjà été utilisée en systématique. Une analyse utilisant un peu moins de 1 kb de séquence codante pour une quinzaine d'espèces de Pinaceae a montré que le gène évolue plus rapidement que les gènes chloroplastiques de la maturase K (*mat*K) et de la sous-unité 5 du NADH-déshydrogénase (*nad*5) (Wang et al.,

2000). Comme l'appartenance à une famille multigénique était possible, des précautions avaient été prises pour éviter des problèmes de paralogie. On a prouvé que tous les clones des différentes espèces d'un genre formaient des groupes monophylétiques. Malheureusement, les clones d'une espèce ne se groupaient pas toujours aussi fortement et étaient parfois entrelacés avec ceux d'autres espèces du même genre. Ces résultats ont été interprétés comme des duplications récentes et suggèrent la prudence lors de toute analyse utilisant des gènes nucléaires.

Granule-bound-starch-synthase

Le gène nucléaire *GBSSI* ou *WAXY* code pour l'enzyme dite de « granule-bound starch synthase » de 59 kDa, dont la séquence codante mesure près de 3 kb. Pour beaucoup d'espèces, il semble que le gène contienne 12 introns et qu'il soit en copie unique dans le génome (Figure 1.2). Dans les Rosaceae cependant, Evans et al. (2000) ont détecté deux copies du gène chez certains individus, ce qui ouvre la voie aux problèmes de paralogie dans des analyses phylogénétiques. Une analyse du gène *GBSSI* entre des espèces du genre *Ipomea* (Convolvulaceae), procurait beaucoup moins de résolution que l'analyse des séquences d'*ITS,* quoique les deux gènes étaient suffisamment variables (Miller et al., 1999). *GBSSI* a aussi été utilisé pour évaluer les relations phylogénétiques au niveau familial. Dans les Poaceae, ce gène affichait un niveau adéquat de variabilité et son analyse procurait beaucoup de résolution, mais, ici encore, peu de support (Mason-Gamer et al., 1998). Les séquences d'exons pourraient être utiles à un bon éventail de niveaux taxonomiques, cependant, le nombre élevé d'introns de taille considérable dans *GBSSI* pourrait nuire à l'amplification et le rendre inutile pour des études à des niveaux familiaux.

Cycloidea

Le gène *CYCLOIDEA (CYC)* code pour un facteur de transcription dont la fonction semble restreinte à la détermination de la forme de la fleur. Il dirige le développement des fleurs zygomorphes chez *Antirrhinum* (Luo et al., 1996), en

collaboration avec au moins six autres gènes. Au moins un de ces gènes, *DICHOTOMA (DIC)*, présente une certaine similarité de séquence avec *CYC*. Il a été démontré que dans la famille des Scrophulariaceae, il existe au moins cinq paralogues de près de 1 kb dans cette famille multigénique, *CYC1A, CYC1B, CYC2, CYC3, CYC4* (Vieira et al., 1999) qui n'auraient qu'un seul exon (Figure 1.2). Dans la famille des Gesneriaceae, l'on retrouve trois de ces gènes, *GCYC1A, GCYC1B et GCYC2* (Citerne et al., 2000). L'orthologie des copies entre les membres de ces deux familles d'angiospermes n'a pas été démontrée, et il se pourrait que les duplications dans la famille multigénique soient courantes et se produisent indépendamment dans plusieurs lignées (Citerne et al., 2000), compliquant de ce fait la détermination des orthologues et des paralogues.

L'intérêt suscité par des gènes responsables du développement de la zygomorphie a été immédiat. En effet, la zygomorphie florale est un caractère d'une grande importance pour toute la biologie des plantes, pour la pollinisation et la classification entre autre. L'existence de multiples copies de *CYCLOIDEA* pourrait compliquer d'éventuelles études de biologie moléculaire et de phylogénies. Tout de même, il semble que lorsque l'échantillonnage est suffisant, les séquences des paralogues sont assez différentes pour être reconnues comme telles (Vieira et al., 1999). Par contre, le taux de variation des orthologues, calculé sur quelques 10 espèces, serait trop faible à l'intérieur de la famille des Scrophulariaceae pour fournir suffisamment de caractères phylogénétiques (Vieira et al., 1999). Pour 13 espèces de Gesneriaceae, *GCYC1* (A et B) présentait un nombre suffisant de substitutions et suggérait des relations congruentes à celles de l'*ITS* (Citerne et al., 2000). Pourtant, le lien tant espéré avec la morphologie florale n'a pu être établi de façon claire. Les orthologues des gènes *CYC* pourraient donc être utiles phylogénétiquement, mais on n'a pas encore déterminé à quels niveaux, ni quels sont les problèmes particuliers qui s'appliquent.

Glutamine synthétases

Les gènes codants pour l'enzyme glutamine synthétase (*ncpGS*) sont importants pour le métabolisme de l'azote et font partie d'une famille multigénique.

Chez *Pisum sativum*, il n'y a qu'une copie de glutamine synthétase chloroplastique (*ncpGS* ou *GS2*), mais ce n'est pas le cas pour toutes les espèces de légumineuses (Emshwiller et Doyle, 1999). Cette enzyme est codée dans le noyau, puis transportée dans les chloroplastes. Le noyau code aussi pour diverses autres glutamine synthétases du cytosol (GS1, GS3A/GS3B, Walker et al., (1995)). La divergence entre les gènes glutamine synthétases date certainement d'avant la séparation des eudicotylédones et des monocotylédones, peut-être même encore plus tôt (Emshwiller et Doyle, 1999). La longueur totale de la région codante est d'environ 1100 pb et la longueur des 11 introns est très variable bien que leur position soit conservée (Figure 1.2). Peu d'études de systématique végétale ont utilisé les séquences de *ncpGS* comme source de données. Des séquences partielles du gène ont été comparées entre 8 espèces d'*Oxalis* (Oxalidaceae) et ont montré que *ncpGS* est variable (3,9 % de sites variables dans les régions codantes et quatre fois plus dans les introns), et que le signal phylogénétique qu'il génère est congruent avec *ITS* (Emshwiller et Doyle, 1999). Du polymorphisme a été détecté chez quelques individus, sans doute attribuable à de l'hétérozygotie. Cela suggère que *ncpGS* puisse être utile à un très fin niveau taxonomique, pour étudier les origines polyploïdes par exemple.

Phosphoglucose isomérase

Le gène phosphoglucose isomérase (*PgiC)* est impliqué dans la biosynthèse du saccharose dans le cytosol et est exprimé continuellement . Chez *Arabidopsis*, la séquence complète fait près de 5 kb et inclut 21 introns (Kawabe et al., 2000). PgiC serait en copie unique, sauf chez certaines espèces de *Clarkia* (Onagraceae) (Gottlieb et Ford, 1997; Ford et Gottlieb, 1999). Dans ce genre, seules certaines espèces expriment les deux copies (Gottlieb et Ford, 1997). Ce gène est légèrement polymorphe chez beaucoup d'organismes dont des invertébrés, des insectes et des plantes, où les substitutions non-synonymes sont souvent anormalement élevées (Liu et al., 1999). Plusieurs études sur le gène *PgiC* dans des espèces de Brassicaceae ont montré que le polymorphisme est attribuable à un haut taux de recombinaison (Liu et al., 1999) et à la sélection pour les hétérozygotes (*balancing selection*) (Kawabe et al., 2000). D'autre part, les

études de *PgiC* dans *Leavenworthia* (Filatov et Charlesworth, 1999; Liu et al., 1999) ont démontré que les différents allèles sont plus « âgés » que les espèces, ce qui en ferait un bon exemple de triage de lignées (*lineage sorting*). Ce gène est donc plus qu'intéressant en ce qui a trait à l'évolution moléculaire, mais le polymorphisme commande la prudence lors des analyses phylogénétiques à de petites échelles taxonomiques.

Alcool déshydrogénase

L'alcool déshydrogénase (*ADH*) est une enzyme glycolitique du métabolisme anaérobique dont l'expression dépend des stress physiologiques. Le gène fait partie d'une petite famille multigénique d'environ trois membres, mais jusqu'à sept chez *Pinus banksiana* (Perry et Furnier, 1996). Il semble y avoir de la confusion quant à la nomenclature des différentes copies : les différents gènes nommés *ADH3* mentionnés dans la littérature (Charlesworth et al., 1998; Koch et al., 2000) ne sont pas homologues alors que *ADH2* de l'orge et *ADH3* du maïs le sont (Gaut et al., 1999; Lin et al., 2001) ! Le gène totalise 1100 nucléotides codants interrompus par neuf introns ou moins (Figure 1.2). Règle générale, la position et le nombre des introns sont conservés, mais plusieurs pertes d'introns sont survenues. De plus, le nombre de copies du gène est variable, même entre espèces proches parentes. Par exemple, *Arabidopsis thaliana* n'a qu'une seule copie de l'*ADH* alors qu'*Arabis*, un genre de la même tribu, en a trois (Koch et al., 2000). Certaines duplications seraient donc récentes : une duplication de l'*Adh2* s'est produite lors de l'évolution du genre *Arabis* mais pas dans les genres apparentés. On nomme maintenant ces deux paralogues *ADH2-1* et *ADH2-2* (Koch et al., 2000). On a aussi découvert des loci totalement dépourvus d'intron dans *Arabis blepharophylla, A. hirsuta* et *A. procurrens* (Koch et al., 2000) ainsi que dans toutes les espèces de *Leavonworthia* échantillonnées (Charlesworth et al., 1998). La perte de tous les introns ne serait pas une synapomorphie, elle se serait produite plus d'une fois dans l'évolution de la famille des Brassicaceae.

L'évolution du gène *ADH* a été intensivement étudiée dans plusieurs organismes et à différentes échelles taxonomiques. Un grand nombre d'études

ont porté sur le polymorphisme de l'*ADH* dans les populations de *Drosophila* et chez les plantes. La recombinaison ou la conversion génique entre les copies de l'*ADH* a été décelée dans plusieurs espèces (Innan et al., 1996; Small et Wendel, 2000); par contre, ces mécanismes ne seraient qu'épisodiques dans l'évolution du gène et semblent aisément détectables. *Arabidopsis thaliana* et *A. lyrata* (Savolainen et al., 2000) sont polymorphes pour l'*ADH*, montrant en moyenne moins de 1% de variation (Miyashita, 2001). Les séquences d'*ADH1A* et *ADH2* ont aussi été utiles pour démontrer l'hybridation dans un complexe hybride de *Paeonia* (Sang et Zhang, 1999) puisqu'on a pu retrouver les allèles des deux parents (vivant ou éteint) au sein de l'hybride. Entre les espèces, il y avait entre 0,1 à 1,6 % de divergence à l'intérieur de chaque loci, surtout concentré dans les régions non-codantes. Par ailleurs, une analyse phylogénétique à une plus grande échelle taxonomique, à l'intérieur de la famille des Brassicaceae, a pu confirmer certaines hypothèses d'évolution, par exemple, la polyphylie du genre *Arabis* (Koch et al., 2000). Beaucoup de séquences de l'*ADH* sont connues et il a été calculé que son évolution ne suit pas une horloge moléculaire stricte, ni entre les lignées de graminées, ni entre les différentes copies (Gaut et al., 1999). Le gène *ADH* semble particulièrement utile à l'étude des populations; cependant, son utilisation à des niveaux taxonomiques supérieurs pourrait être risquée, bien que faisable.

Chalcone synthétase

La famille multigénique des gènes chalcone synthétase *(CHS)* a été largement étudiée du point de vue de l'évolution moléculaire mais peu utilisée en systématique végétale. L'enzyme chalcone synthétase est impliquée dans les premières étapes de la biosynthèse des flavonoïdes et dans plusieurs autres processus. Contrairement à l'évolution d'autres familles multigéniques, le nombre de copies de *CHS* ne semble pas être soumis à des restrictions (Clegg et al., 1997). Dans le génome de *Petunia hybrida*, huit gènes *CHS* sont connus; six chez *Ipomea purpurea*; huit chez la fève; une seule chez les Brassicaceae (Koch et al., 2000). Le gène est habituellement long de près de 2000 pb codantes interrompues de deux introns (Figure 1.2) (Helariutta et al., 1996; Huttley et al.,

1997; Koch et al., 2001). L'évolution de cette famille multigénique est caractérisée par des duplications multiples dans des lignées indépendantes. Plutôt que de devenir des pseudogènes, les gènes dupliqués sont rapidement recrutés pour accomplir de nouvelles fonctions. Le taux de substitution non-synonyme y est alors beaucoup plus élevé que pour une comparaison de deux séquences transcrites. Dans certains cas, les substitutions non-synonymes étaient plus élevées que les mutations dans les régions non-codantes (Huttley et al., 1997)! L'enzyme stilbène synthétase par exemple, a évolué indépendamment de plusieurs lignées des chalcone synthétases (Durbin et al., 2000). D'autres exemples existent dans *Gerbera*, où une enzyme avec une nouvelle fonction (encore inconnue) a évolué à partir d'un gène de chalcone synthétase (Helariutta et al., 1996). Les gènes dupliqués peuvent aussi demeurer actifs en tant que chalcone synthétase et créer de la redondance. Dans ce cas, les multiples copies montrent souvent des patrons d'expression différents (Durbin et al., 2000). Probablement à cause de cette redondance partielle, les différentes copies de chalcone synthétases ne sont pas soumises à une pression de sélection élevée et les substitutions entre les orthologues sont nombreuses. Ce phénomène important pourrait nuire à l'utilisation des gènes *CHS* en systématique moléculaire. On a en effet remarqué que de la conversion génique ou de la recombinaison entre locus peut survenir (Huttley et al., 1997; Koch et al., 2000). L'étendue du phénomène n'a pas été évaluée, ce qui compromet d'autant plus l'interprétation d'arbre phylogénétique des gènes *CHS*. Les études de polymorphisme à l'intérieur du locus *CHS-A* chez *Ipomea purpurea* ont montré que la variabilité est grande entre les dix allèles de *CHS-A*, surpassant celle des allèles de l'*ADH1* (Huttley et al., 1997).

Le gène *CHS* a, d'autre part, été utilisé pour déterminer les relations phylogénétiques à un niveau taxonomique supérieur, entre une quarantaine d'espèces de Brassicaceae (Koch et al., 2001). Les sites variables comptaient pour 43,9 % des sites et la majorité était des changements non-synonymes, et les branches basales de l'arbre phylogénétique résultant de l'analyse étaient bien supportées. De mineures différences de topologie ont été notées entre les arbres

de *CHS* et ceux de *matK*, une séquence chloroplastique. Le gène chalcone synthétase possède donc un potentiel certain en systématique moléculaire des plantes; il pourrait être utile à des niveaux taxonomiques allant de infraspécifique à intergénérique. On devrait procéder avec prudence à l'heure de l'inférence à l'évolution des espèces d'après les analyses avec *CHS*, surtout à cause d'un niveau intermédiaire de conversion génique que peuvent subir les allèles ou loci (Sanderson et Doyle, 1994).

Les gènes nucléaires utilisés en systématique moléculaire représentent bien la variabilité et la plasticité du génome nucléaire, en ce qui a trait à la fonction, à la structure et à l'évolution moléculaire. Ils codent pour des protéines impliquées dans des processus physiologiques ou du développement, d'autres sont des protéines de structure ou excrétées. Certains gènes ont un taux de substitution élevé, d'autres sont très conservés. Le nombre d'introns varie grandement d'un gène à l'autre et certains en sont complètement dépourvus. Une constatation importante est que les gènes en copie strictement unique dans tous les groupes de plantes sont extrêmement rares. Des gènes que l'on croyait en unique copie sont plutôt présents en deux copies dans certains groupes de plantes. *LEAFY* est l'un des seuls gènes pour lequel on n'a pas trouvé plus d'une copie dans un diploïde, quoique peu d'individus aient été échantillonnés jusqu'ici. Nous avons vu que les copies d'une famille multigénique sont soumises à des mécanismes d'évolution particuliers. Ces paralogues peuvent subir de la recombinaison ou de la conversion génique entre eux, ils peuvent devenir des pseudogènes ou encore ils peuvent être redondants fonctionnellement. Ce sont tous des mécanismes qui ont un impact lorsque ces gènes sont utilisés en systématique moléculaire. Les gènes en copie unique présentent donc un avantage certain pour la reconstruction phylogénétique. Il est primordial d'insister sur le caractère unique de l'évolution et de l'utilité de chacun des gènes nucléaires; à cause de la grande variabilité du génome nucléaire, ils ne peuvent être réduits à aucune généralité.

Tableau 1.1 Liste partielle des gènes nucléaires couramment utilisés en systématique moléculaire des plantes. Les niveaux taxonomiques approximatifs où les séquences codantes (les exons) des gènes seraient utiles sont indiqués, ainsi que les études d'où proviennent les données. Le pourcentage de variation signifie le pourcentage de sites variables dans l'alignement.

Nom (abréviation)	**Niveaux taxonomiques**	**Organismes étudiés**	**Taux de variation (en %)**	**Source**
ARN polymérase II (*RPB2*)	Supra familial	Eucaryotes Plantes vertes Gentianales	 43,9	Sidow et Thomas, 1994 Denton et al., 1998 Oxelman and Bremer, 2000
LEAFY (*LFY*)	Supra familial	Angiospermes Angiospermes		Frohlich et Meyerowitz, 1997 Frohlich et Parker, 2000
glycéraldéhyde 3-phosphate déshydrogénase (*gapA*) (cp)	Familial	Berberidaceae		Adachi et al., 1995
glycéraldéhyde 3-phosphate déshydrogénase (*gapC*)	Supra familial	Angiospermes		Martin et al., 1993
Arginine décarboxylase (*ADC*)	Familial	Brassicaceae	34,3	Galloway et al., 1998
Légumines (*leg*)		Ranunculaceae		Lang et Fisher, 1995
Vicilines	Générique à Familial	Genre *Lens* Sterculiaceae	 42	Saenz de Miera et Pérez de la Vega, 1998 Whitlock et Baum, 1999
Phytochromes, (*PHY A à PHY E*)	Tribal à supra	Angiospermes PHYA/C et PHY B/E		Mathews et al., 1995, Donoghue et Mathews, 1998; Mathews et Donoghue, 1999

	familial	Milletieae (Leguminosae) PHYA, PHYE	30,8	Lavin et al., 1998
		Poaceae PHYB	51,6	Mathews et al., 2000
		Celastraceae PHYB	48,5	Simmons et al., 2001
4-coumarate : coenzyme A ligase (*4CL*)	Familial	Pinaceae		Wang et al., 2000
granule-bound starch synthase (*GBSSI*)	Familial	Poaceae	51,1	Mason-Gamer et al., 1998
		Convolvulaceae		Miller et al., 1999
		Rosaceae		Evans et al., 2000
CYCLOIDEA (*CYC*)	Familial	Gesneriaceae		Citerne et al., 2000
		Scrophulariaceae		Vieira et al., 1999
glutamine synthétase (*ncpGS*)	Spécifique à générique	genre *Oxalis* (Oxalidaceae)	3,8	Emshwiller et Doyle, 1999
phosphoglucose isomérase (*PgiC1; PgiC2*)	Générique	genre *Clarkia* (Onagraceae)		Gottlieb et Ford, 1997; Ford et Gottlieb, 1999
		Leavenworthia stylosa		Filatov et Charlesworth, 1999
		genre *Leavenworthia*	7	Liu et al., 1999
		genre *Arabidopsis*		Kawabe et al., 2000
alcool déshydrogénase (*ADH*)	Générique à familial	*Arabidopsis thaliana*	7,9	Innan et al., 1996
		genre *Gossypium*	2,7	Small et al., 1998; Small et Wendel, 2000
		Poaceae		Gaut et al., 1999
		Paeonia (Paeoniaceae)		Sang et Zhang, 1999
		Brassicaceae	37,3	Koch et al., 2000
		genre *Arabidopsis*		Charlesworth et al., 1998; Savolainen et al., 2000; Miyashita, 2001
Chalcone synthétase (*CHS*)	Générique à familial	*Ipomea* (Convolvulaceae)	4,9	Huttley et al., 1997; Durbin et al., 2000
		Brassicaceae	40,7	Koch et al., 2000; Koch et al., 2001

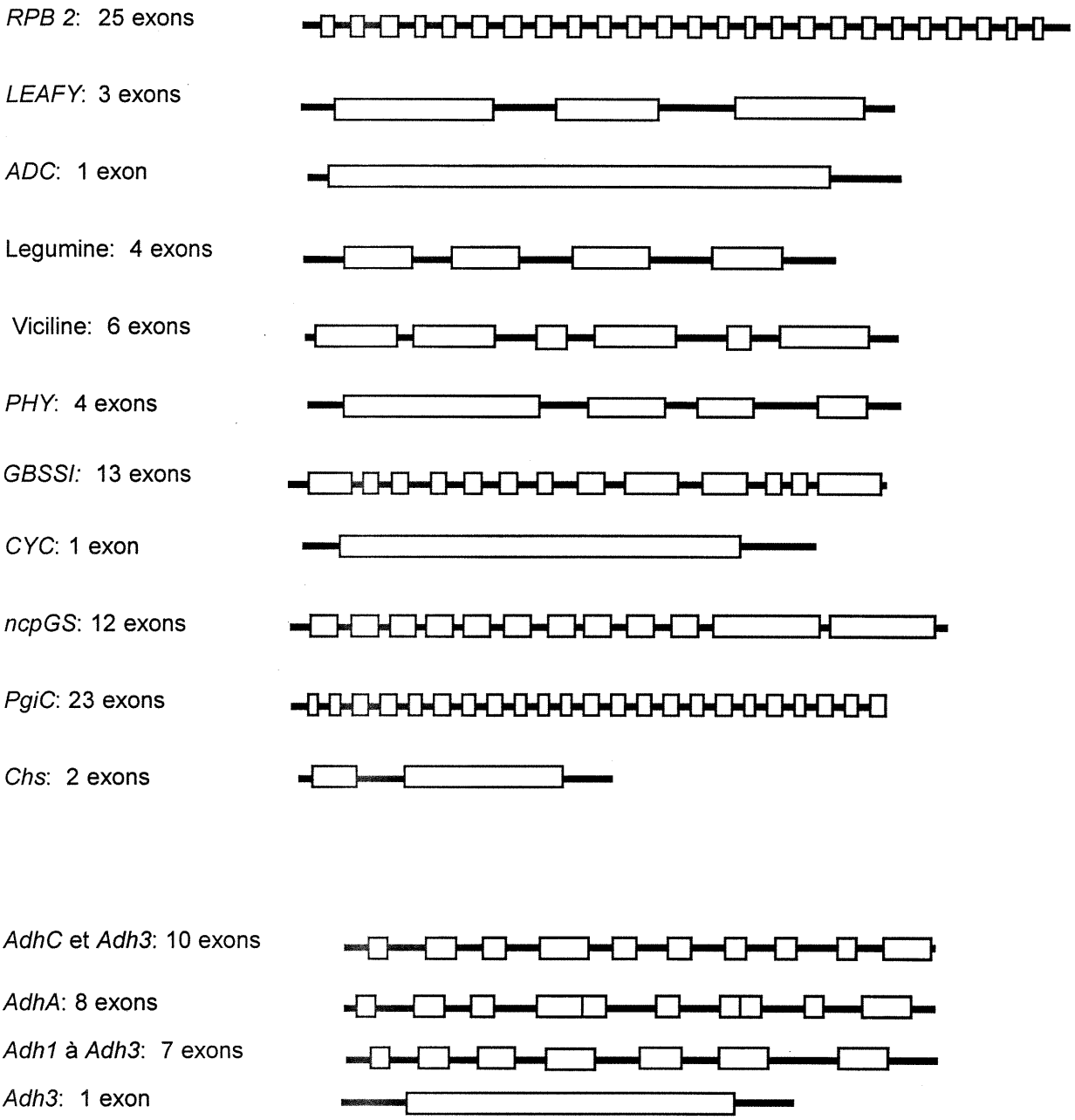


Figure 1.2: Position des introns pour quelques-uns des gènes nucléaires utilisés en systématique moléculaire. Les exons sont représentés par des boîtes. L'information est tirée de Denton et al. (1998), Frohlich et Meyerowitz (1997), Frohlich et Parker (2000), Galloway et al. (1998), Genbank, Whitlock et Baum (1998), Mathews et al. (1995), Evans et al. (2000) , Mason-Gamer et la. (1998), Citerne et al. (2000), Eshmwiller et Doyle (1999), Ford et Gottlieb (1999), Kawabe et al. (2000), Small et Wendel (2000), Koch et al. (2000), Charlesworth et al. (1998), Huttley et al. (1997), Koch et al. (2000). La taille des images n'est pas proportionnelle à la taille réelle des gènes.

Le gain ou la perte d'introns comme caractère phylogénétique?

Les trois génomes des végétaux subissent des changements génomiques à grande échelle, appelés changements génomiques rares ou, plus communément, mutations structurelles, qui peuvent, tout comme les substitutions, fournir des caractères utiles en systématique moléculaire (Rokas et Holland, 2000). Ces mutations pourraient rassembler des groupes d'espèces ou identifier des orthologues dans une famille multigénique. Parmi des exemples de changement rares, l'on retrouve le gain ou la perte d'introns, l'intégration de rétrotransposons, les séquences signatures, le changement d'ordre de gène des génomes chloroplastiques ou mitochondriaux, les duplications géniques, la variation dans le code génétique (Rokas et Holland, 2000).

Les résultats de mes recherches montrent qu'au moins trois espèces d'un genre possèdent un nouvel intron dans le gène *LEAFY*. C'est pourquoi je m'intéresserai plus particulièrement à la présence d'introns comme marqueur phylogénétique. Un intron est une séquence non codante interrompant la séquence codante d'un gène, mais retirée au moment de la maturation de l'ARN messager (épissée). Plusieurs types d'introns existent, certains peuvent s'auto-épisser d'autres non. Dans le génome nucléaire l'on ne retrouve que ceux de type épissable (*spliceosomal*), ils nécessitent un complexe appareillage d'ARN catalytique et de protéines, le complexe d'épissage (*spliceosome*), pour être épissés. La distribution phylogénétique des introns d'un gène, ainsi que leur absence de mécanisme d'auto-épissage ou de transposition suggère que les insertions ou les pertes d'introns sont des événements rares. Si ce sont des événements uniques, la présence ou l'absence de certains d'entre eux pourrait être utilisée comme marqueur phylogénétique, pour reconnaître des gènes orthologues ou pour marquer des groupes d'espèces lorsqu'on compare des gènes orthologues dans plusieurs espèces.

Afin de connaître la valeur d'un caractère comme la présence ou l'absence d'un intron dans un contexte phylogénétique, il faut aborder la structure de l'intron,

le mécanisme d'épissage et les fonctions des introns. Je discuterai aussi brièvement des différentes hypothèses émises pour expliquer l'apparition des introns dans le génome des eucaryotes.

Le mécanisme d'épissage des introns nucléaires

La découverte de la présence des introns à l'intérieur d'un gène en 1977 a étonné, et les recherches ont rapidement élucidé les mécanismes d'épissage des introns. Dans un intron de type nucléaire, trois régions sont essentielles pour son l'épissage : les sites de clivage localisés en 3' et en 5' de l'intron et la région d'embranchement (*branch point*, Figure 1.3) située à environ 18 à 60 pb en amont du site 3' de l'intron (Brown et al., 1996). On représente les sites conservés du site de clivage par GT...AG (Figure 1.3): l'intron débute presque toujours avec un GT et se termine par AG (Simpson et Filipowicz, 1996). La séquence des exons aux abords du site de clivage est moins conservée, mais dans près de la moitié des cas, elle est AG/intron/GT. La séquence de la région d'embranchement est YUNAN et est semblable dans tous les eucaryotes (Brown et al., 1996), quoiqu'elle montre une grande variabilité.

Quatre étapes de maturation d'un ARN messager se produisent avant son entrée dans le cytoplasme. Dans le noyau, 1) le gène est d'abord transcrit par l'ARN polymérase, puis 2) il subit des modifications à ses extrémités : ajout de la queue de polyA et de la coiffe 5' ; 3) le pré-ARNm est alors soumis à l'épissage, c'est à dire que les introns sont excisés et 4) les extrémités d'exons sont réunies. La réaction d'épissage du pré-ARNm est catalysée par le complexe d'épissage (*spliceosome*), un grand complexe ribonucléoprotéique dynamique, impliquant plus de 100 protéines et 50 pARNn (petits ARN nucléaires, *small nuclear RNA*). Les pARNn sont associés à des protéines et sont alors appelés les pRNPn (petites ribonucléoprotéines nucléaires, *small nuclear ribonucléoprotéins*) qui s'assemblent et se désassemblent continuellement lors de la réaction d'épissage. L'une des six pRNPn impliquées dans l'épissage peut reconnaître la séquence consensus du site de clivage en 5' de l'intron (Lewin, 2000). L'intron est retiré par deux étapes de transestérifications. Tout d'abord, une attaque de l'adénosine de

la région d'embranchement (le A de la séquence YUNAN) clive le site 5' de l'intron, formant un lasso et laissant libre l'extrémité de l'exon en amont de l'intron (Figure 1.3 A). Le site 3' de l'intron est ensuite clivé par une attaque de l'extrémité libre de l'exon libéré à l'étape précédente (Figure 1.3 B). C'est ensuite qu'a lieu la ligation des exons et le relâchement du lasso (Figure 1.3 C) (Simpson et Filipowicz, 1996; Lodish et al., 1999). Certaines étapes de l'épissage nécessitent de l'ATP, c'est donc un processus énergivore (Lodish et al., 1999). Bref, pour que la réaction d'épissage se produise sans erreurs, la reconnaissance du site en 5' est primordiale, de même que l'est celle de l'adénosine du site d'embranchement et la formation du lasso.

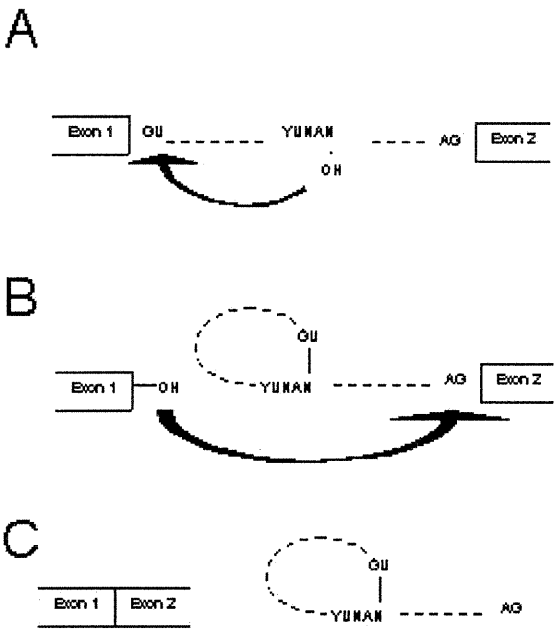


Figure 1.3 Deux transestérifications (A et B) et la réunion des exons (C) mènent à l'épissage des introns. Plus tard, l'intron sera linéarisé.

Les mécanismes d'épissage des introns nucléaires sont bien caractérisés pour beaucoup d'organismes modèles eucaryotes. Ces mécanismes sont souvent semblables; pourtant lorsque des introns de levure ou de mammifères sont incorporés dans une plante, ils ne sont pas toujours reconnus (Simpson et Filipowicz, 1996). Il existe donc des différences importantes entre ces règnes. L'un des signes distinctifs des introns végétaux est qu'ils sont plus riches en AT, mais surtout en T, que les introns animaux ou de levure. La présence de 20% de plus de AT dans l'intron que dans les exons, tout juste en amont du site de clivage 3', est requise pour que l'épissage se produise correctement (Brown, 1996). Les séquences d'introns nucléaires, même s'ils ne sont pas codants, ne sont donc pas

libres de toute pression de sélection puisque la séquence de certaines régions doit être conservée afin que l'intron soit correctement excisé.

Il existe d'autres types d'introns qui ne nécessitent pas de complexe nucléoprotéique (complexe d'épissage) ni d'ATP pour être épissés et peuvent eux-mêmes catalyser la réaction d'épissage par leur ARN, sans l'aide d'aucune autre molécule : ils sont autoépissés. Trois groupe de ces introns existent, les groupe I, I et III, présents surtout, mais pas uniquement, dans les génomes d'organites cellulaires (Buchanan et al., 2000; Lewin, 2000). Ces introns sont cependant beaucoup moins communs que les introns nucléaires. Les réactions de l'épissage sont semblables à celles de l'épissage d'un intron nucléaire, elles procèdent aussi par deux transestérifications (Lewin, 2000), à la différence qu'elles sont catalysées par l'ARN de l'intron lui-même.

Les introns de groupe II se trouvent surtout dans les gènes codant pour les protéines des génomes mitochondriaux et chloroplastiques. La séquence des introns de groupe II n'est pas nécessairement conservée, mais ils se replient selon une structure secondaire de tige-boucle qui est cruciale pour la réaction d'épissage (Lewin, 2000). La structure de ces tiges et boucles ressemble d'ailleurs aux pRNPn du complexe d'épissage. En fait, malgré que l'auto-épissage par l'ARN des introns de groupe II soit possible *in vitro*, l'action d'une maturase codée à l'intérieur même de l'intron et traduite en protéine est nécessaire pour que l'épissage soit rapide *in vivo*. La maturase sert à stabiliser la structure tridimensionnelle du pré-ARNm, elle a aussi une activité de rétrotranscription qui permet la rétrotransposition des introns de groupe II ailleurs dans le génome (Lewin, 2000). Les similarités entre les introns de groupe II et les introns nucléaires ont d'ailleurs servi de point de départ à l'une des hypothèses avancées pour expliquer la présence des introns dans le génome nucléaire, qui est la théories des introns tardifs (Lodish et al., 1999).

La présence d'introns de groupe II dans des gènes mitochondriaux a déjà été utilisée comme caractère phylogénétique, pour cerner des groupes

d'angiospermes (Joly et al., 2001) par exemple. La conclusion de l'étude montre que malgré que l'intron de COXII.i3 soit auto-épissable et mobile, sa présence peut servir de marqueur phylogénétique pour certains groupes, malgré des pertes multiples.

Origines et fonctions des introns

La découverte des introns et la plus grande connaissance des mécanismes d'épissage nous amène immédiatement aux questions de l'origine et de la fonction des introns dans le génome des eucaryotes.

Une origine récente ou ancienne des introns

Deux hypothèses principales peuvent expliquer l'origine des introns dans le génome. L'une prédit que les introns seraient arrivés très tôt dans l'évolution de la vie, l'autre, qu'ils sont apparus beaucoup plus tard, et les tenants de l'une ou l'autre hypothèse ne peuvent parvenir à un consensus. La ressemblance entre les mécanismes d'épissage des introns nucléaires et de ceux de groupe II a servi à l'élaboration de l'une des hypothèses. Il a été proposé que les séquences d'ARN des pRNPn du complexe d'épissage, qui interagissent avec les sites d'épissage en 5' et en 3' et entre elles, soient fonctionnellement analogues aux tiges et boucles des introns de groupe II (Lodish et al., 1999; Lewin, 2000). Cette hypothèse explique alors la présence des introns dans le génome nucléaire par l'invasion d'introns de groupe II provenant des organites suivie de la perte de leur structure interne et donc de leur capacité d'auto-épissage. L'invasion aurait été permise grâce à la capacité de rétrotransposition de la maturase. Pour appuyer cette hypothèse, on note que des introns de groupe II auxquels on a enlevé certains domaines de la séquence agissent comme des introns nucléaires (Lodish et al., 1999). Les introns auraient pu s'insérer dans des endroits favorables du génome, appelé sites de proto-clivage (*protosplice site)* (Dibb et Newman, 1989). Cette théorie, qui postule que les introns sont apparus récemment dans l'histoire des eucaryotes, est appelée *intron late theory*, la théorie des introns tardifs (Cavalier-Smith, 1985; Rogers, 1985; Palmer et Logdson, 1991).

Les hypothèses alternatives d'évolution des introns nucléaires proposent que les introns soient apparus beaucoup plus tôt dans l'évolution du génome nucléaire des eucaryotes (Li, 1997). L'une des hypothèses postule qu'au départ, dans les tous premiers eucaryotes, les introns étaient des séquences non codantes présentes entre de courts domaines codant pour de petites protéines. Puis, au fil de l'évolution, plusieurs domaines juxtaposés se sont intégrés en un seul gène et les séquences intervenantes sont devenues des régions transcrites épissées : des introns. On l'appelle la théorie exonique des gènes (*exon theory of genes)* (Gilbert et al., 1986). Certains proposent alternativement que des introns étaient présents dans l'ancêtre commun des eucaryotes et des procaryotes, ces introns étaient auto épissés au départ, mais ont ensuite perdu cette capacité. Selon cette hypothèse, les introns ont été perdus dans les lignées menant aux procaryotes et chez certains protistes, qui n'ont presque aucun intron de type nucléaire dans leur génome. Pendant ce temps, les introns nucléaires ont occupé de plus en plus d'importance dans le génome des animaux, mycètes et plantes. C'est l'hypothèse des introns hâtifs (*intron early theory*).

Les gains et les pertes d'introns sont possibles

Afin d'appuyer l'une ou l'autre des hypothèses, une variable d'importance capitale nous manque encore : c'est la facilité relative de la perte ou du gain d'un intron. Effectivement, s'il était établi avec certitude que les pertes d'introns sont très courantes alors que les gains sont rares, la théorie des introns hâtifs pourrait être considérée plus crédible que celle des introns tardifs. Par contre, si les pertes d'introns sont rares alors que les gains sont fréquents, s'insérant à un même site de proto-clivage par exemple, c'est la théorie des introns tardifs qui serait favorisée, et la présence d'un intron pourrait aisément constituer un caractère homoplasique. Les connaissances actuelles de l'évolution moléculaire du génome nucléaire, empiriques ou théoriques, ne nous permettent pas de choisir entre ces deux hypothèses. Il est fort probable qu'aucune ne peut expliquer la distribution de tous les introns, et qu'une combinaison de mécanismes sont impliqués dans l'évolution d'un gène (Tyshenko et Walker, 1997). Pourtant, quiconque voulant

utiliser la présence des introns en un locus donné comme marqueur phylogénétique devrait connaître la fréquence relative des pertes et des gains. Cette information est nécessaire afin de tenir compte de l'homoplasie dans l'évolution du caractère.

La perte des introns d'un gène se produit entre autre lorsqu'il y a transcription inverse d'un ARN messager, qui est suivie par une recombinaison homologue où l'ADN sans intron remplace la séquence existante (Baltimore, 1985). Les acquisitions d'introns peuvent quant à elles survenir lorsque des éléments transposables s'insèrent dans le génome et qu'ils sont ensuite parfaitement épissés (Giroux et al., 1994), ou lorsqu'une portion de séquence codante contenant un la séquence MAGR (A/C A G A/G) caractéristique d'un site de proto-clivage est dupliquée. La portion se trouvant entre ces deux sites serait automatiquement correctement épissée (Venkatesh et al., 1999).

D'après des observations empiriques tirées de la littérature, on peut considérer sous certaines réserves que la présence d'un intron à la même position dans des gènes homologues serait une apomorphie alors que l'absence d'un tel intron peut fort possiblement être un caractère homoplasique. Il existe en fait de nombreux exemples de pertes indépendantes. Par exemple, alors que la majorité des angiospermes ont neuf introns dans le gène *ADH*, toujours positionnés aux même endroits, l'*ADHA* du coton diploïde (*Gossypium*) en a sept. Les copies de l'*ADH* chez les Brassicaceae ne possèdent pas ces deux même introns, qui auraient été perdus indépendamment et cette double absence ne serait pas due à une origine commune (Small et Wendel, 2000). De plus, la copie *ADH3* du genre *Leavenworthia* (Charlesworth et al., 1998), ainsi que l'une des copies de *Arabis procurrens, A. hirsuta* et *A. blepharophylla* (Koch et al., 2000) sont caractérisées par l'absence totale d'introns. La perte de tous les introns se serait produite deux fois dans l'évolution des Brassicaceae Dans les gènes de catalase des angiospermes, on a montré que des pertes indépendantes de différents ensembles d'introns font en sorte que l'absence d'un intron à une position donnée dans le gène est souvent un caractère homoplasique (Frugoli et al., 1998).

Les pertes et gains d'introns ont été utilisés avec succès dans une perspective phylogénétique. Chez les vertébrés par exemple, la présence d'un intron dans le gène *RAG1b* caractérise l'un des grands groupes de poisson (Venkatesh et al., 1999). Chez les plantes, l'une des trois copies du gène catalase (la copie *CatA*) aurait acquis un intron en une nouvelle position chez toutes les espèces d'*Oryza* échantillonnées (Frugoli et al., 1998; Iwamoto et al., 1998). Le survol des séquences de la triose phosphate isomérase (*Tpi*) d'organismes de plusieurs règnes a mis en évidence l'acquisition de sept introns au cours de l'évolution de ces lignées (Logsdon et al., 1995). À ma connaissance, aucune acquisition indépendante d'introns en un même site n'a été démontrée. Pourtant, l'hypothèse des sites de proto-clivage rend cette éventualité probable. En effet, selon cette hypothèse, un nouvel intron ne s'insérera pas complètement au hasard, mais à l'intérieur d'une séquence consensus (Dibb et Newman, 1989). L'absence d'exemples d'acquisition indépendante d'introns pourrait être due à un manque d'échantillonnage plutôt qu'à l'improbabilité de cet événement.

Les introns ont-ils une fonction?

La possibilité de créer des plantes transgéniques et d'étudier l'expression et la traduction des gènes a rendu plus facile l'étude de la fonction des introns que l'étude de leur évolution. Grâce à ces outils, on a pu découvrir que la présence d'un intron dans un gène peut affecter son expression, dans la majorité des cas en l'accroissant (Simpson et Filipowicz, 1996). C'est dans les monocotylédones que le phénomène a été le plus étudié et où l'accroissement de l'expression est le plus drastique. Par exemple, la présence d'un intron de l'*ADH1* du maïs peut augmenter l'expression d'un gène rapporteur de 100 fois ! Le même effet est observé chez les dicotylédones, avec une augmentation de 2 à 5 fois seulement. Cette conséquence de la présence des introns est modulée par plusieurs facteurs, entre autre par la position de l'intron dans le gène, par le promoteur associé ou par l'état physiologique des cellules (Simpson et Filipowicz, 1996). Il ne semble pas que la hausse d'expression du gène soit due à une séquence signal de l'intron puisque de larges délétions des introns peuvent être effectuées sans

compromettre l'effet de l'augmentation de l'expression. La hausse de l'expression d'un gène grâce à la présence d'un intron peut être partiellement expliquée par la plus grande stabilité de l'ARNm. La présence des introns aurait aussi fortement influencé l'évolution des protéines au cours de l'évolution moléculaire (Li, 1997), la recombinaison dans les introns aurait permis l'échange d'exons (*exon shuffling*), créant ainsi de nouvelles protéines.

On ne s'entend donc pas sur les causes de l'apparition des introns dans le génome nucléaire des eucaryotes. Ont-ils été insérés à partir d'éléments transposables et, si c'est le cas, depuis quand ? Ou ont-ils seulement été intégrés dans un gène ? Ce manque de connaissance pose un problème du point de vue de la qualité comme marqueur phylogénétique de la présence d'un intron puisqu'on ne peut pas proposer de probabilité de gain d'un intron. En fait, on ne peut s'assurer de l'homologie des caractères de présence ou absence d'un intron. La présence d'un intron dans un gène codant peut accroître de beaucoup son expression et représente donc un avantage. Mais pourtant, il existe des exemples clairs de perte d'intron nucléaire, et même de pertes indépendantes du même intron.

Conclusion

Tout au long de ce chapitre, les particularités du génome nucléaire des végétaux et leur influence lors des analyses phylogénétiques ont été abordées. Comparativement aux génomes chloroplastiques et mitochondriaux, le génome nucléaire évolue beaucoup plus rapidement et le taux de substitution est extrêmement variable de région en région. La structure du génome nucléaire est aussi beaucoup plus complexe que celle des autres génomes, caractérisée entre autre par un grand nombre de séquences non-codantes, souvent répétitives ou mobiles, par l'omniprésence des familles multigéniques et la rareté des gènes en copie unique. Ces gènes dupliqués n'évoluent pas nécessairement indépendamment, ce qui les rend difficiles d'utilisation pour des reconstructions phylogénétiques. Pour les systématiciens moléculaires, l'idéal est d'utiliser un gène en copie unique. Par contre, considérant que le nombre de copies d'un

gène varie d'un groupe taxonomique à l'autre et qu'il est très laborieux méthodologiquement de prouver l'existence d'une seule copie d'un gène dans un génome, toute étude visant à évaluer l'utilité d'une nouvelle région en systématique doit tenir compte de tous ces problèmes potentiels.

La présente étude vise l'évaluation du gène *LEAFY* pour la systématique des légumineuses et sera décrite au second chapitre. *LEAFY* est un gène en copie unique chez les diploïdes étudiés jusqu'ici et dont l'évolution est moyennement rapide. La séquence en acides aminés est comparable entre les Embryophytes. Son utilité sera testée pour la systématique des légumineuses, mais surtout pour l'une des sous-familles : les Caesalpinioideae.

Par ailleurs, les gènes nucléaires sont aussi caractérisés par la présence d'introns épissables. Ces introns ne sont pas mobiles, pourtant il existe quelques exemples d'insertion ou de disparition d'introns survenues au cours de l'évolution de certaines lignées. Peu d'études se sont penchées sur l'intérêt phylogénétique de la présence d'un intron dans une région donnée du génome. Celles qui existent révèlent que les acquisitions indépendantes sont rares alors que les pertes multiples semblent plus courantes. Comme très peu de données existent sur ce sujet, on ne peut être certain de la valeur de la présence d'un intron dans un gène pour caractériser l'évolution des espèces. Le séquençage du gène *LEAFY* dans les légumineuses a mis en évidence la présence d'un nouvel intron chez certaines espèces d'un genre de la sous-famille des Caesalpinioideae. Cet intron pourrait être un nouveau caractère réunissant les espèces porteuses. D'ailleurs, l'examen préliminaire de la distribution phylogénétique de cet intron est présenté au second chapitre.

Chapitre 2 :

Phylogenetic utility of the coding sequences of the *LEAFY/FLORICAULA* gene and of a new intron discovered in this gene in the Caesalpinioideae

Résumé

La présente étude montre que le gène *LEAFY* pourrait être utile pour la systématique des Caesalpinioideae et des légumineuses, quoique le signal phylogénétique qu'il génère est différent de celui de l'intron du gène chloroplastique *trn*L. Nous avons cloné et séquencé une région de *LEAFY* allant de l'extrémité 3' du second exon jusqu'à l'extrémité 3' du troisième exon pour 36 espèces provenant des trois sous-familles de Leguminosae, mais avec une insistance particulière dans les Caesalpinioideae, la plus basale des sous-familles. Le taux de substitution des exons de *LEAFY* fournit de l'information phylogénétique au niveau taxonomique de la famille et à des niveaux inférieurs. Le second intron est plus variable et sa séquence ne peut être comparée entre des espèces de différentes tribus ou sous-familles. *LEAFY* est en copie unique pour la majorité des espèces étudiées ici et les exceptions pourraient être des polyploïdes. Les trois sous-familles des légumineuses sont bien reconnues dans les analyses des séquences en acides nucléiques de *LEAFY*, par contre, de l'incongruence est trouvée à l'intérieur de la tribu Detarieae entre *LEAFY* et l'intron *trn*L. L'incongruence entre *LEAFY* et *trn*L reste inexpliquée, quoique plusieurs scénarios d'explication soient abordés. Certaines des espèces de *Brownea* (Detarieae) échantillonnées possèdent un nouvel intron qui pourrait être un bon caractère taxonomique.

Abstract

The present study shows that the nuclear gene *LEAFY* is useful for the systematics of the Caesalpinioideae and the Leguminosae. However, the phylogenetic signal generated by *LEAFY* is different that of the chloroplast *trn*L

intron. We cloned and sequenced a partial sequence of *LEAFY*, from the 3' end of the second intron to the 3' end of the third exon, for species from the three Leguminosae subfamilies, but with a higher density in the Caesalpinioideae, the more basal subfamily. The rate of substitution in *LEAFY* gives a phylogenetic signal at and below the family level. The second intron is highly variable, and the sequence cannot be compared between species of different tribes, or subfamilies. *LEAFY* is a single copy gene in the Leguminosae, except for two species that may be polyploids. The three subfamilies are well recognised by the phylogenetic analyses of DNA sequences. However, incongruence between *trn*L and *LEAFY* trees is found in the tribe Detarieae. This incongruence remains poorly explained, but several explanations are explored. Some species of the genus *Brownea* (Detarieae) sampled possess a new intron that appears to be a useful taxonomic character.

Introduction

Very few low-copy nuclear sequences are currently utilised for phylogenetic analyses in spite of the innumerable possibilities offered by this genome. In contrast to chloroplast genes, which tend to show homogenous rates of evolution, the rate of substitution in nuclear genes shows a greater variability (Wolfe et al., 1989). This greater variability may provide a source of characters important for understanding species evolution at broad ranges of taxonomic levels. The largest databases of sequences used for phylogenies of plants are nonetheless from the chloroplast genome, and to a lesser extent from nuclear rDNA. Reasons are numerous and justified: chloroplast sequences are easier to obtain and the interpretation of the results is more straightforward. Serious difficulties, practical and analytical, limit the general use of nuclear DNA regions in phylogenetic studies. Common difficulties with nuclear genes are that the number of copies of a gene per genome varies commonly among species (Doyle, 1992; Doyle and Davis, 1998; Oxelman and Bremer, 2000; Small and Wendel, 2000), recombination and gene conversion may occur between members of a multigene family (Moniz de Sa and Drouin, 1996; Ford and Gottlieb, 1999; Gaut et al., 1999)

and, at a low taxonomic level or when working with recent speciation events, persistence of divergent alleles of highly variable genes in populations or species may occur (Filatov and Charlesworth, 1999; Koch et al., 2000). In this study, we evaluate the usefulness of the *LEAFY* gene, a single copy nuclear gene from plants, as a molecular marker in the legume family, with a special interest in subfamily Caesalpinioideae.

LEAFY is a single copy nuclear gene useful in phylogenetic studies of angiosperms and gymnosperms species (Frohlich and Meyerowitz, 1997; Frohlich and Parker, 2000). LEAFY controls determination or indetermination of meristems in various species. This gene is a transcription factor that was first characterized in *Antirrhinum majus (*Scrophulariaceae) where a loss-of-function of *FLORICAULA* prevents the plant from making the transition from inflorescence to flower (Coen et al., 1990). In the model species *Arabidopsis thaliana* (Brassicaceae), where the genetic mechanisms of floral promotion and development are now well understood, *LEAFY* is involved in two signalling pathways that promote flowering (Devlin and Kay, 2000). Gibberellins phytohormones activate *LEAFY* in a day-length independent pathway, and the CONSTANS (CO) transcription factor indirectly regulates *LEAFY* in the photoperiodic pathway (Devlin and Kay, 2000). These two independent pathways act on different regions of the *LEAFY* promoter (Blazquez and Weigel, 2000). LEAFY, in turn, directly activates floral meristem identity genes of the ABC model in cooperation with other proteins such as UFO (Parcy et al., 1998). In *Arabidopsis*, *LEAFY* is expressed only in the floral meristem thus providing a determinate fate to the meristem. In numerous other species like peas (Leguminosae) and *Jonopsidium acaule* (Brassicaceae), *LEAFY* orthologs have the same function of promoting floral development (Singer et al., 1999; Shu et al., 2000), thus allowing determination of the meristem (Theissen, 2000). However, in some other species, *LEAFY* function is dramatically different. In *Impatiens balsamina*, *IMPFLO* expression and loss-of-function mutants do not show any role in floral meristem establishment (Souer et al., 1998).

In other organs or other species, in rice and pea leaves and *Jonopsidium acaule* bracts for example, *LEAFY* orthologs (termed *RFL*, *UNIFOLIATA*. and *vcLEAFY* respectively), maintain vegetative meristems in an indeterminate state (Shu et al., 2000; Theissen, 2000). In loss-of-function mutants of *UNIFOLIATA* in *Pisum sativum*, compound leaves become single leaflets (Hofer et al., 1997; Hofer and Ellis, 1998; Gourlay et al., 2000), which, with the two stipules, look like trifoliolate leaves.

The first cloning of *FLORICAULA* in *Antirrhinum* revealed that it was a novel class of protein (Coen et al., 1990). A transcription factor function had been proposed for this protein based on a proline-rich and an acidic region at the N-terminal end of the protein sequence. Parcy et al. (1998) later demonstrated that LEAFY localizes to the nucleus, binds DNA in a sequence-specific manner, and can mediate transcription activation in yeast, thus showing the transcriptional activation function of the LEAFY protein. The acidic region is present in *Antirrhinum* and *Arabidopsis*, but in gymnosperms, this region is replaced by a basic region. Despite its highly divergent sequence and structure, expression of *NEEDLY* (the *LEAFY* ortholog from *Pinus)* can rescue severe *lfy* phenotypes in *Arabidopsis*, indicating that NEEDLY can accomplish the same critical functions as LEAFY in plant development (Theissen, 2000). Thus, the nucleotide sequence of a transcription factor may not be the most important feature for its role in flower development. Because nothing is known about the three-dimensional structure, the active domains and the interactions with other proteins, we still do not know how the LEAFY protein works.

In all angiosperms investigated to date, *LEAFY* has been found in a single copy for diploids and in two copies for tetraploids (Shu et al., 2000; Theissen, 2000). *LEAFY* has also been found in *Pinus*, *Welwitschia* and *Gnetum*, and two copies of *LEAFY* occur in *Pinus* (Frohlich and Meyerowitz, 1997). A *LEAFY* phylogenetic study across spermatophytes has served as a basis for elaborating a novel theory of flower origin in the angiosperms, in addition to demonstrating the

phylogenetic utility of *LEAFY* amino acid sequences at high taxonomic levels (Frohlich and Parker, 2000).

The focus of the present study is the subfamily Caesalpinioideae of the Leguminosae. The Leguminosae are the third largest flowering plant family, comprising about 18000 species grouped in 650 genera. A large amount of diversity is observed in this economically important family in term of habitat, morphology and biochemistry. Many of these aspects have been studied in diverse species and are summarised in the series Advances in Legume Systematics (part 1 to part 9). A close relationship of the legumes with a group of families including the Polygalaceae, not expected based on traditional classifications (Dickinson, 1981), was repeatedly evidenced by the molecular analyses from chloroplast sequences (Doyle, 1995; Bremer et al., 1998; Doyle et al., 2000).

Traditionally, three subfamilies were recognized in the Leguminosae: the basal paraphyletic Caesalpinioideae, the highly specialized Mimosoideae, and the mostly temperate and economically important Papilionoideae. Despite numerous studies in the systematics of the Leguminosae, the relationships for the most basal lineages and among subfamilies are not clearly understood. The most basal lineages of the family are grouped together in the subfamily Caesalpinioideae (Doyle, 1995; Doyle et al., 2000; Bruneau et al., 2001). This subfamily is further subdivided into four tribes (previously five), namely Cercideae, Cassieae, Caesalpinieae and Detarieae s. l. In recent systematic studies, however, Cassieae and Caesalpinieae were shown to be polyphyletic (Tucker et Douglas, 1994; Doyle, 1995). The largest tribe of the Caesalpinioideae is the Detarieae s. l., which is monophyletic if including Amherstieae (or Macrolobieae) and where the floral morphology is exceptionally diversified (Polhill et al., 1981; Bruneau et al., 2000). Mimosoideae were shown to be derived from groups of Caesalpinieae (Doyle et al., 2000; Luckow et al., 2000; Bruneau et al., 2001). Papilionoideae appears as a monophyletic group, sister to a clade containing members of the Cassieae and the Caesalpinieae, and the Mimosoideae (Bruneau et al., 2001).

This large family is also very old: the first fossil data (pollen and wood) are detected as early as the Maastrichtian (65-70 My of years (Polhill et al., 1981; Raven and Polhill, 1981; Herendeen et al., 1992)). The first lineages evidenced belong to the tribes Detarieae and Cassieae in the Caesalpinioideae, and the three subfamilies are well established by the Lower Eocene (38-54 My of years (Herendeen et al., 1992)). Although they are not the oldest fossils discovered, molecular studies suggest that a monophyletic Cercideae is sister to the remainder of the family (Doyle et al., 2000). The oldest lineages would therefore have retained their basic characters for 60 My.

The systematics of the Leguminosae is mostly from chloroplast genes, *rbc*L (Doyle, 1995; Doyle et al., 1997; Doyle et al., 2000), *trn*L (Bruneau et al., 2000; Bruneau et al., 2001), and *mat*K (Lavin et al., 2001). However, there is interest in confirming chloroplast results with nuclear sequences for relationships at the infrafamilial level. The nuclear ribosomal ITS sequences (Käss and Wink, 1997), the histone H3-D locus (Doyle et al., 1996a), and the nuclear phytochromes loci (PHYA, PHYA1 and PHYE) (Lavin et al. 1998) are the only nuclear loci that have been used in the Leguminosae and they are restricted to the tribal or generic level in the subfamily Papilionoideae.

Considering the potential problems and usefulness for phylogenetic studies of nuclear genes in general and of the *LEAFY* gene in particular, we asked whether this gene could be used for resolving species relationships in the Leguminosae. The general objective of this study is to assess the reliability of the phylogenetic signal of *LEAFY* in the Leguminosae. The current study was designed to explore *LEAFY* sequence evolution over a wide range of taxonomic levels and divergence times. Our study sought to determine the taxonomic level at which *LEAFY* is maximally informative for phylogeny reconstruction, and to compare the rates of substitution and level of utility of *LEAFY* to those of a chloroplast gene, the *trn*L intron, studied in the same group of Leguminosae. Specifically, we will consider the number of copies of *LEAFY* in the nuclear

genome of the Leguminosae, the extent of sequence variation at the family level and below, and whether the phylogenetic signal from *LEAFY* is congruent with that from a chloroplast gene.

Material and Methods

Taxon sampling

We designed a sampling strategy that covers different levels of taxonomic diversity to determine at which level a phylogenetic signal can be detected (Table 2.1). We sampled among the three Leguminosae subfamilies: Caesalpinioideae, Mimosoideae and Papilionoideae. Within the Caesalpinioideae, we sampled taxa from the four tribes of the subfamily (Cassieae, Caesalpinieae, Cercideae, Detarieae). Within the tribe Detarieae, we sampled a large number of genera (22/61) ensuring that we sampled among most clades resolved in the chloroplast *trn*L analysis of Bruneau et al. (2000, 2001). In order to detect if phylogenetic signal is present at the generic level, we also sequenced four *Brownea* and two *Browneopsis* species (*Brownea* group, Detarieae).

Table 2.1 : Species of Leguminosae sequenced for partial *LEAFY* sequences and the published *trn*L intron sequences (Bruneau et al., 2000; Bruneau et al., 2001). The second intron length of *LEAFY* is shown in the last column. Taxonomy follows Polhill (1994); generic groups and subtribes are given in the first column. Dashes mean that the information is not known yet for this particular individual.

	Species	accession and information voucher [a]	Genbank accession number *trn*L	*LEAFY*	Intron 2 length (bp)
Caesalpinioideae					
Detarieae					
Amherstia	*Amherstia nobilis* Wall.	Breteler 13507 (WAG)	-		220
	Amherstia nobilis Wall.	Baker 490 (KEP)	AF365210	-	-
	Humboltia laurifolia Valh	Rickson sn (OSC)	AF365211		215-216
	Tamarindus indica L.	JBM 213876	AF365206		361-371
Berlinia	*Berlinia confusa* Hoyle	Breteler 13373 (WAG)	AF365215		>1008
	Englerodendron usambarense Harms	Herendeen -17-XII-97-2 (US)	AF365218		155
Brachystegia	*Brachystegia laurentii* (De Wild.) Louis ex Hoyle	Wieringa 2925 (WAG)	AF365251		206
	Brachystegia mildbraedii Harms	J. de Wilde 11718 (WAG)	AF365254		206
Brownea	*Brownea capitella* Jacq.	Breteler 13506 (WAG)	AF365196		-
	Brownea coccinea Jacq.	Baker 600 (MT)	AF365195		218 and 308
	Brownea grandiceps Jacq.	Klitgaard 621 (K)	-		
	Brownea grandiceps Jacq.	Klitgaard 67015 (AAU)	AF365193		215-217
	Brownea leucantha Jacq.	Klitgaard 666 (K)	AF365197		216-217
	Brownea multijuga Britton & Killep.	Klitgaard 67001 (AAU)	AF365194		215-216
	Browneopsis disepala (Little) Klitgaard	Klitgaard 67032 (AAU)	AF365198		-
	Browneopsis ucayalina Huber	Klitgaard 684 (K)	AF365199		-
Crudia	*Crudia gabonensis* Pierre ex Harms	Wieringa 2585 (WAG)	AF365172		222
	Gossweilerodendron balsamifera (Verm.) Harms	Wieringa 3233 (WAG)	AF365166		220-221

	Guibourtia demeusii (Harms) J. Léonard	Wieringa 2396 (WAG)	AF365175	155
	Guibourtia pellegriniana J. Léonard	van Bergen 425 (WAG)	AF365176	-
	Oxystigma manii Baill.	Bruneau 1057 (K)	AF365167	-
Cynometra	*Cynometra ramiflora* Miq.	no 1987-1990 DuPuy 123 (K)	AF365117	206-207
	Schotia afra (L.) Thumb. Ou Thunb.	Hodgkiss 1 (BOL)	AF365122	235
	Scorodophleus zenkeri Harms	Breteler 14073 (WAG)	AF365125	220-218
	Umtizia listeriana T. Sim	Schrire 2602 (K)	AF365126	348
	Zenkerella citrina Taub	Cheek 7614 (K)	AF365127	-
Hymenaea	*Hymenaea verrucosa* Gaertn.	Herendeen 11-XII-97-3 (US)	AF365162	-
Hymenostegia	*Afzelia bella* Harms	Chase (K)	AF365128	117
	Hymenostegia ngounyensis Pellegr.	Wieringa 2579 (WAG)	AF365142	222
	Plagiosiphon emarginatus (Hutch. & Dalz) Léon.	Breteler 13354 (WAG)	AF365153	223-223
	Saraca dives Pierre	Manos 1419 (DUKE)	AF365158	94
Macrolobium	*Anthonotha macrophylla*	Bruneau 1059	-	>1063
	Anthonotha macrophylla P. Beauv.	Wieringa 2996 (WAG)	AF365234	-
	Macrolobium ischnocalyx Harms	Klitgaard 669 (K)	AF365201	155
Cercideae				
Bauhiniineae	*Bauhinia bohniana* L. Chen	Douglas 766 (MEL)	AF365056	515
Cassieae				
Cassiinae	*Chamaecrista sp.*	Klitgaard 654 (K)	AF365093	185
	Senna alata (L.) Roxb.	Bruneau 1076 (K)	AF365091	154
Dialiinae	*Dialium guianensis* (Aubl.) Sandw.	Klitgaard 686 (K)	AF365079	558
	Dialium guineense Willd.	Breteler 14748 (WAG)	AF365081	
Caesalpinieae				
Peltophorum	*Delonix elata* (L.) Gamble	Herendeen 20-XII-97-1 (US)	AF365106	302-305
	Schizolobium parahyba (Vell.) Blake	Klitgaard 694 (K)	AF365108	282
Caesalpinia	*Caesalpinia decapetala* (Roth) Alston	Herendeen 19-XII-97-1 (US)	-	307
	Caesalpinia calycina Benth.	Lewis 1885 (K)	AF365064 -	-
Mimosoideae				
	Entada polyphylla Benth.	Klitgaard 613 (K)	-	297-300

Acacia caven (Molina) Molina	JBM 386-89	AF365041	455
Inga sp.	Klitgaard 677 (K)	AF365046	-
Papilionoideae			
Apios americana Medik.	Joly CTSP (JBM)	-	505
Pisum sativum L.		-	AF035163565
Dussia tessmanni Harms	Klitgaard 673 (K)	-	500
Polygalaceae			
Polygala comosa (Schkuhr)	Wieringa 3462 (WAG)	AF365036	-

[a] The following botanical gardens are listed and abbreviated : Arnold Arboretum, Boston, USA; Montréal Botanical Garden, Canada, JBM; Royal Botanical Gardens Kew, UK, Royal Botanical Melbourne, Australia; Wageningen Botanic Garden, Netherlands.

Molecular methodology

Total DNA was isolated from 0,3 g silica gel-dried leaves using a modified CTAB extraction protocol (Doyle and Doyle, 1987), or for a few samples, using a DNeasy QIAGEN extraction kit (Qiagen Inc, Missaussauga, Ontario Canada). *LEAFY* was amplified from the second to the third exon (Fig 2.1). PCR conditions were in a 50 µl final volume as follows: 0.5 units Taq polymerase, 1X buffer (Roche Molecular Diagnostic), 200 µM each dNTP, 0.4 to 0.5 µM both primers, 5% DMSO, 1 µl DNA. Typical conditions for PCR were 4 minutes of initial denaturation at 94 C followed by 45 cycles of denaturation at 94 °C for 30 seconds, annealing at 60 °C for 1 minute, and extension at 72 °C for 20 seconds to 3 minutes, ended by a final extension of 7 minutes at 72 °C. Using set of primers from the literature (LFsxl3, LFR1, Frohlich and Meyerowitz (1997)) we first amplified the region from the second to the third exon for a few genomic DNA. It was then possible to design new internal primers (LFsxl3C LFsxl3D and LFR1C and LFR1D, Figure 2.1, table 2.2). For several samples, a nested PCR was performed, using LFsxl3C and LFR1C as the first set of primers and LFsxl3D and LFR1D as the second set. Amplification reactions were purified using QIAquick PCR purification kit (Qiagen Inc, Missaussauga, Ontario Canada).

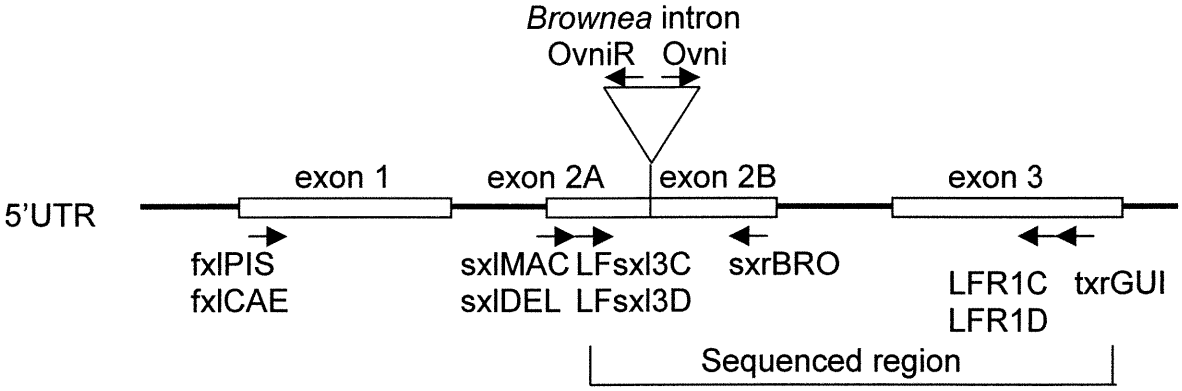


Figure 2.1 : Position of introns and exons designed in this study in the *LEAFY* gene of the Leguminosae, showing new intron on *Brownea* and primers relative position. Exon 1 was not sequenced.

Because of low specificity of LFsxl3C or LFsxl3D and LFR1C or LFR1D primers (they often could amplify very efficiently unidentified DNA regions), we designed a new primer downstream at the limit of the third exon. We used the Genome Walker approach, this protocol (Clontech Laboratories, Inc, Protocol #PT3042-1, Version # PR03300) is designed to find unknown genomic sequences adjacent to known sequences. Following that protocol, we digested genomic DNA from *Guibourtia demeusii* (Detarieae) with six restriction enzymes: *Apa*I, *Dra*I, *EcoR*V, *Pvu*II, *Stu*I, *Bsty*I, and then ligated digested DNA with the adaptor. Using adaptor primers and gene specific primers, we performed nested PCR to find the sequence downstream from our known sequence. We then gel-purified the bands of interest with the Qiagen gel extraction kit (Qiagen Inc, Missaussauga, Ontario Canada) and either sequenced them directly or cloned them. This procedure allowed us to design the txrGUI primer.

Table 2.2: Primers designed in this study to amplify *LEAFY* from the Leguminosae.

Gene region	Primer	Orientation	Sequence (5' - 3')
Exon 1	fxlPIS	direct	ATGGATCCAGACGCATTCACCGCCAGCTTGTTYAARTGGGA
	fxlCAE	direct	ATGGATCCHGATGCMTTCACWGCMWSVYTKTTYAARTGGGA
Exon 2	LFsxl3B	direct	GTGGCRCGKGGSAARAAGAAKGG
	LFsxl3C	direct	GCGGGGSAARAAGAAGGGCCTYGA
	LFsxl3D	direct	GCCTYGACTACCTCTTCCATC
	sxlDEL	direct	GGGGGATGTGAGAGRCARAGAGARCAYCCATTCAT
	sxlMAC	direct	GAGGGATGTGAGAGACAGAGAGAACACCCATTCAT
	sxrLEG	reverse	CGATGTTCTGGACTTGGATYAAGAAYTCACSGCRTTG
Exon 3	LFL3B	direct	CGGACATIAATAAGCCIAARATGCGICAYTA
	LFR1B	reverse	AGCTGRCGGAGYYKGTRGGSAC
	LFR1C	reverse	TGRCGGAGYBGGTGGGSACRTACC
	LFR1D	reverse	ACRTACCATATKGAAAGGCG
	LFR2B	reverse	GGACGTGICGIARIYKIGTIGGIACRTACC
	LFtxrB	reverse	CGYARTGTCGCATYTTBGGCT
	txrGUI	reverse	GCACTATTCCTCTCGGCGTGACAAAGCTGACGG
Brownea intron	OVNI	direct	CCCRTGTTTTATGTCATCAAGTCCTCAAGTTTAT
	OVNIR	reverse	TGAAGACGCAGCAGCAAAGTAAMTGACATC

A few amplifications of the first exon were successful using the fxlPIS and sxrLEG primers, using the same reaction conditions as described above, but with an annealing temperature of 63 °C. Those sequences, although incomplete, allowed us to design the sxlMAC and the sxlDEL primers (Fig 2.1).

Taq-amplified products were cloned either in a pGEM –5Zf(+) vector (Promega Corporation, Madison WI, USA) or the pCR4 TOPO vector (Invitrogen Corporation, Baltimore, Maryland). The pGEM-5Zf(+) was made in our laboratory by cutting the multiple cloning site with *EcoR*V followed by the addition of dTTP with Taq polymerase (Marchunk et al., 1990). Ligation was performed as recommended in the Promega technical manual No.042 with the modifications noted by Forest and Bruneau (2000). Following the ligation, we heat-shocked XL1-MRF' (Stratagene, La Jolla, Californie) chemically competent cell (Sambrook et al., 1989). Transformants were plated on LB agar with blue/white screening ability and appropriate antibiotics (tetracycline and ampiciline). pCR4 TOPO was used as recommended in the TOPO TA cloning kit protocol (Invitrogen, Kit: K4575-01, Protocol: 000808 25-0275 TOPO TA Cloning kit for sequencing Version 2). Colonies were screened by PCR (Dallas-Wang et al., 1998), all transformed colonies were picked when transformation efficiency was low (3 to 10 transformed colonies in the plate), and up to 20 were chosen when transformation was very efficient. Selected colonies were incubated overnight in LB broth and appropriate antibiotics. Plasmid DNA was extracted using a standard protocol (Sambrook et al., 1989) or QIAprep Spin Miniprep Kit (Qiagen Inc, Missaussauga, Ontario Canada).

Sequencing

Both strands of templates were sequenced using automated cycle-sequencing method in an ABI 310 sequencer (Applied Biosystems, Foster City, California). Cycle sequencing was performed with the BigDye^{tm} terminator cycle ready reaction kits with 250 ηg plasmid DNA following instructions from the

manufacturer (Applied Biosystems, Foster City, California). Primers for sequencing were from the plasmid vector.

Sequence analysis

Sequences were verified, corrected and complementary strands were assembled with Sequencher^{tm} 3.0 software (Gene Code Corporation, Ann Arbor, Michigan, USA). We always performed a Blast similarity search in GenBank (National Center for Biotechnology Information, 2001b) to verify if sequences corresponded to *LEAFY* gene. Sequences were aligned with ClustalX 1.81 program (Thompson and Jeanmougin, 2000), but alignment was refined by eye.

Parsimony analyses (MP) were performed with PAUP* 4b8 (Swofford, 2000). Gaps were treated as missing. When all taxa were analyzed, only coding sequences were included. An initial series of trees was generated by retaining a maximum of MP trees per replicate of 500 random addition replicate. This initial set of trees was used as starting trees for a heuristic analysis with 1000 heuristic searches with tree bisection-reconnection (TBR) branch swapping and random order of taxon addition. Additional analyses were performed with characters from only the first codon position, only the second position, only the first and second position, and only the third position. Another analysis was done in which all sites where proportionally reweighted based on the minimum RCI (Rescaled Consistency index). A combined analysis of *trn*L and *LEAFY* was also performed. Parsimony analyses and reweighting were repeated until tree length stabilized. A Maximum Likelihood (ML) search of tree was performed under the GTR+I+1 model, where all the parameters are estimated from the data. Trees were rooted with *Bauhinia bohniana*, a member of the tribe Cercideae of the Caesalpinioideae, because Cercideae were shown to be sister to the remainder of the Leguminosae in molecular analyses with *trn*L (Bruneau et al., 2001) and *rbc*L (Doyle, 1995). Support values were calculated by the Jackknife procedure using the fast Jackknife option for 10 000 replicate.

Synonymous and non-synonymous substitutions, as well as the proportion of variable sites were calculated using the MEGA software (Kumar et al., 2001). Synonymous and non-synonymous substitutions were calculated following the method proposed by Li, Wu and Luo (1985). The number of steps per character, consistency index (CI) and retention index (RI) at particular positions or domains were determined using the MacClade 4.0 software (Maddison and Maddison, 2000). The incongruence length difference (Farris et al., 1994) test was also applied between *LEAFY* and *trn*L data using PAUP 4 * b8 with 100 replicates .

Results

The final data set contained 73 new partial *LEAFY* sequences and a published one from 36 species representing 32 genera from the three subfamilies of Leguminosae. Sequences were from the 3' end of the second exon to third exon (Fig. 2.1). Amplification with primers LFsxl3C at the 3' end of exon 2 and LFR1C of exon 3 recovered most often one fragment of variable length. The observed variation in length was due to insertions or deletions in the second intron (Table 2.1). The length of the aligned exon regions sequenced is 114 bp for the second exon and 286 bp for the third exon.

In *Brownea grandiceps*, *B. leucantha* and *B. multijuga* (but not *B. coccinea*), amplification fragments were about 1120 bp. Examination of the sequences showed that an additional 490 bp region intervening the second exon was present. This large insertion is more likely to be an intron than a transposon because none of the sequences showed characteristics typical of transposons (e.g., direct or inverted repeats flanking the insert). The splicing sites are not the typical AG/GT sequence but rather C/AAA, and AYA/AA. When submitted to NetPlantGene, the exon / intron splicing site prediction server (Center for Biological Sequence Analysis, 2001), splicing sites characteristic of a plant intron were identified. The sequence variation observed in this new intron is less than that seen in intron 2 of *LEAFY* (6% of variable sites, compared with 22% in intron 2), but is considerably higher than in the chloroplast in *trn*L intron (1,9%) for the same set of taxa.

In species of *Brownea* possessing this additional intron, a fragment of 600 bp (the approximate length the amplification fragment would be without the additional intron) was sometimes also amplified. However, when gel-cleaned, cloned and sequenced, this band was not a *LEAFY* sequence, but an unidentified DNA region with no homology to *LEAFY* when blasted in GenBank. We surveyed for the presence or absence of this additional intron in close relatives of *Brownea* using primers OVNIR specific to this new intron and sxrBRO or txrGUI, but we were

unable to amplify the new intron from species of *Browneopsis* and *Macrolobium* (Table 2.3).

Table 2.3 : Presence or absence of the intron intervening the exon 2 region of the *LEAFY* gene for a few members of the *Brownea* group and potentially related genera.

Species surveyed	Intron
Brownea coccinea Jacq.	Absent
Brownea grandiceps Jacq.	Present (518 – 519 bp)
Brownea multijuga Britton & Killep.	Present (519 bp)
Brownea leucantha Jacq.	Present (519 – 520 bp)
Browneopsis disepala (Little) Klitgaard	Absent
Browneopsis ucayalina Huber	Absent
Macrolobium ischnocalyx Harms	Absent

In some samples, primers LFsxl3B or LFR1B at the 5' end, and LFsxl3C or LFR1C at the 3'end, amplified very efficiently an unidentified region of DNA unrelated to *LEAFY,* except for the primer regions. This same sequence was obtained for *Zenkerella citrina, Scorodophleus zenkerii, Oxystigma manii* and *Browneopsis disepala,* with only 26 variable sites from a total of 542 nucleotides. We also found a *LEAFY* pseudogene in *Hymenaea verrucosa, Polygala comosa* and *Dialium guianensis.* Pseudogenes were identified by the presence of frame shift mutations and stop codons. For *P. comosa* and *H. verrucosa*, the true *LEAFY* gene was never amplified.

Some species in the *LEAFY* phylogenetic tree occurred in an unexpected position. One of the clones of the Mimosoideae, *Entada polyphylla,* was separated from other Mimosoideae in a monophyletic group with clones of *Afzelia bella*, a member of Detarieae. No nucleotide differs between this *Entada polyphylla* clone and one *Afzelia bella* clone while one nucleotide difference exists between the two *Afzelia* clones. This *Entada polyphylla* clone (clone #1) is therefore the likely result of contamination and we decided to exclude it from further analyses. A different conclusion was drawn for the *Brownea coccinea* clone #8 (Detarieae), which grouped with clones of *Delonix elata* #9, a species of the tribe Caesalpinieae , rather than with other Detarieae. Contamination was also

suspected in this case, but we noted 43 nucleotide differences between the *B. coccinea* #8 and the *D. elata* #9 sequences, while there are only five differences between two *D. elata* clones (#9 and #4). The elevated number of differences between this *B. coccinea* and *D. elata* sequences cannot be easily explained by a contamination and thus, sequence of *B. coccinea* #8 was retained in all subsequent analyses.

Because of the unexpected position of *Tamarindus indica* in the *LEAFY* phylogenetic tree, this species was amplified, cloned, and sequenced a second time using the sxlMAC and sxRGUI primers (clones #7 and #9) instead of LFsxl3C and LFR1C (clones #4, #11 and #12). These new *Tamarindus* clones do not form a monophyletic group with previous *Tamarindus* sequences, but instead are included in a sister clade that includes the former *Tamarindus* clones, as well as numerous other taxa (Fig 2.4). This double check has not been done yet for *Brownea coccinea,* another sequence whose phylogenetic position is doubtful.

Alignment

Alignment in the exon was unambiguous, except at the very 3' end, where sequencing was less clear. Only three indels were found in the coding regions. They were in the third exon region and in triplets of nucleotides. In contrast, there is a high level of variation in length (table 2.1) and in sequence (Fig 2.2) in the intron. Within the Detarieae s. l., alignment of the intron was reliable and characterised by large gaps in many sequences. When all taxa from the three subfamilies of Leguminosae are considered, alignment in the intron region was not reliable, and homology could not be assessed. We therefore chose to discard the intron sequences from the analyses of the complete taxa sampling.

Sequence analyses and statistics.

A 51,8% GC content is found in the *LEAFY* coding sequences (table 2.4) and 39,4% in the noncoding sequences (intron). There is no significant nucleotide compositional variation among taxa (x^2 = 59,436, df = 219, P = 1,000). At the third

codon position only, GC content is slightly higher, 39.4% to 77% depending on the exon analysed (table 2.4). Among the legumes, 60% of the coding sites of *LEAFY* are variable; this proportion falls to 34% when the calculation is made for the Detarieae only (Fig 2.2). In the second intron, more than 90% of the sites are variable for the whole data set, and about 50% among the Detarieae. The extent of variation in *LEAFY* is thus slightly higher than for the *trn*L intron, for which the percentage of variable sites is 39% among the legumes.

Table 2.4: Variation and sequence characteristics for the coding sequence of the *LEAFY* gene among all the Leguminosae.

Region	Statistic	1st position	2nd position	3rd position
Exon 2A	nb variable / total sites	7 / 16	6 / 15	14 / 15
	G+C content	24.4+38.6	17.7+0.4	28.4+48.6
	Ti/Tv	3.3	0	2.3
	CI [a]	0.57	na[b]	0.24
	RI	0.80	na	0.67
Exon 2B	nb variable / total sites	11 / 22	10 / 23	18 / 23
	G+C content	27.3+23.1	22.2+13.4	22.7+47.1
	Ti/Tv	0.4	0.6	2.9
	CI	0.71	1	0.45
	RI	0.95	1	0.78
Exon 3	nb variable / total sites	47 / 96	38 / 95	86 / 95
	G+C content	35.1+18	21.7+21.2	30.9+28.5
	Ti/Tv	1	0.9	1.3
	CI	0.67	0.64	0.40
	RI	0.89	0.92	0.76

[a] = CI and RI were calculated from the 200, 000 most parsimonious trees resulting from the analysis including the three codon positions (399 sites). CI was calculated without uninformative characters

[b] = not applicable

For the whole data set, the overall ratio of substitution/synonymous sites is 0.521 and the ratio of substitution/non-synonymous sites is 0.052. The average ratio of synonymous substitution to non-synonymous substitution is 10.8 (± 7.3 SE). Although the standard error is high, this ratio is always higher than one, indicating that there is no evidence of positive selection acting on the whole *LEAFY* gene.

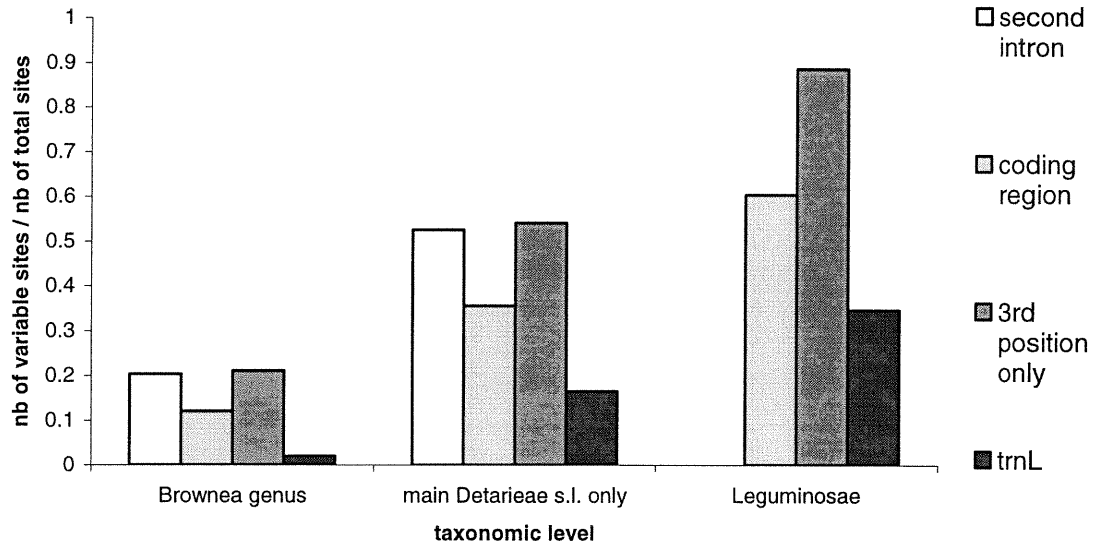


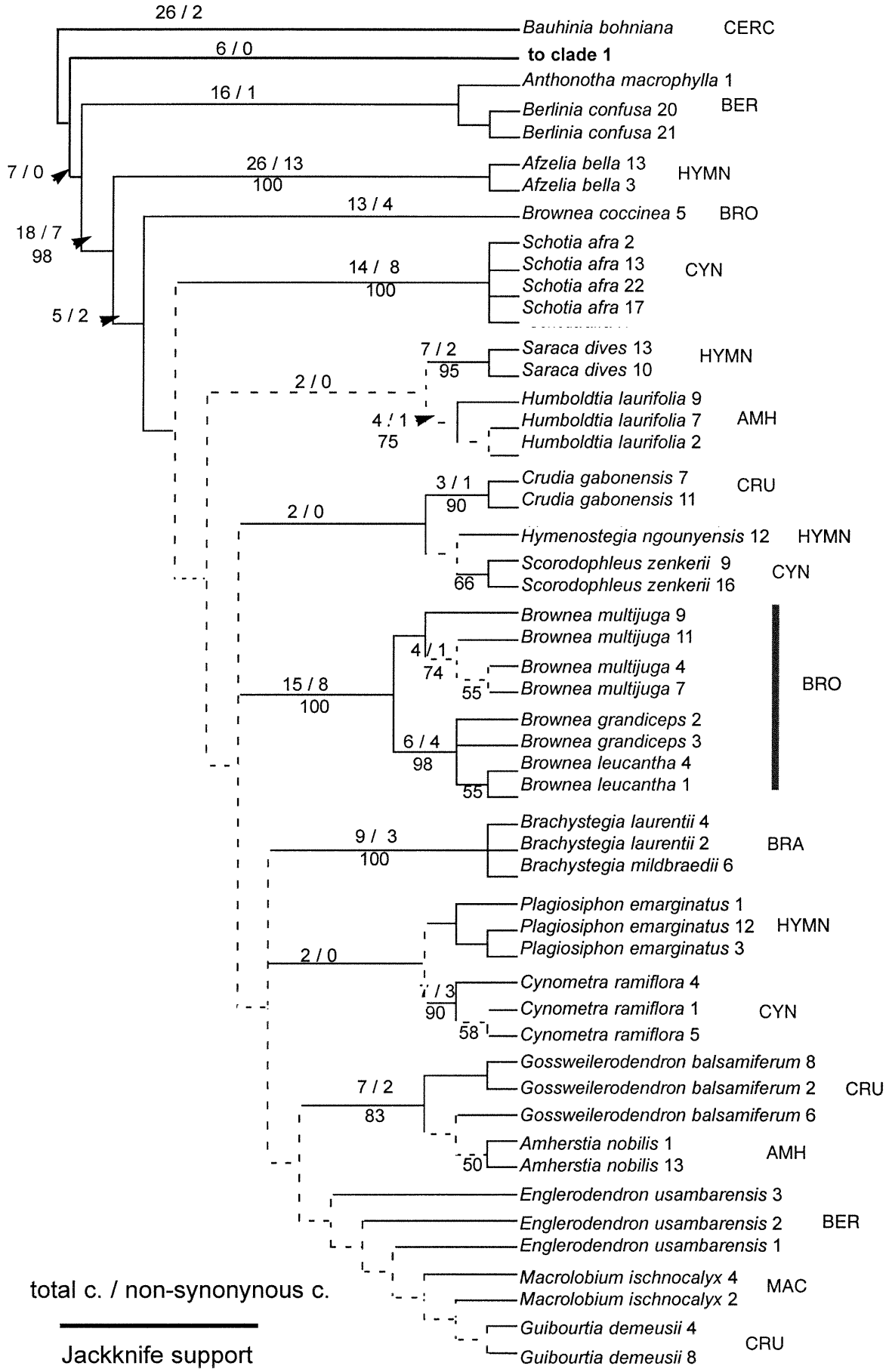
Figure 2.2 . Variation in sequences at various taxonomic levels represented by number of variable sites per number of total sites in different regions of the *LEAFY* gene and in the *trn*L intron.

Phylogenetic analyses

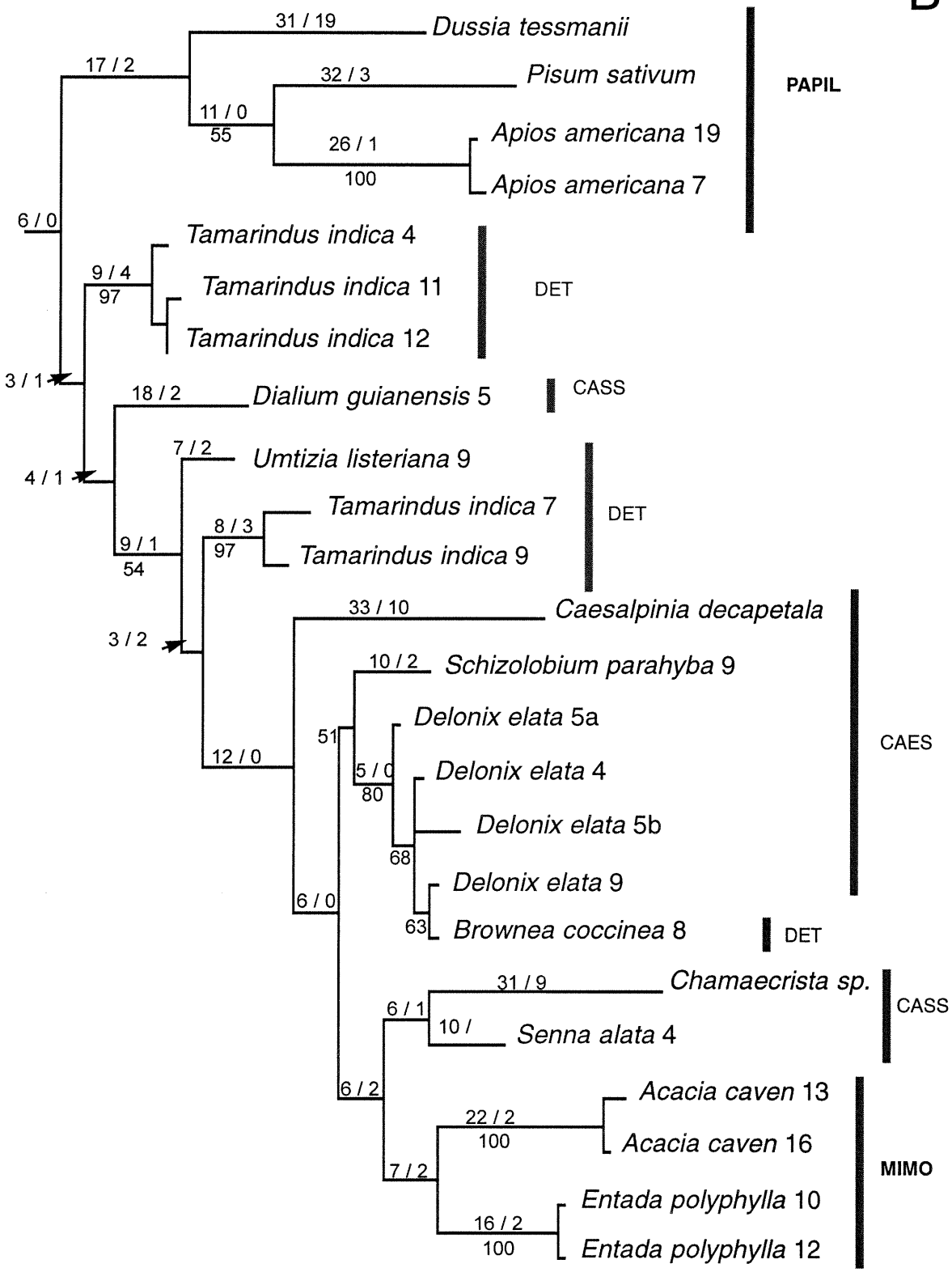
The parsimony analysis of the *LEAFY* gene, including every clone of all the individuals sampled (74 sequences from 36 species), was performed with the coding sequences only. This region provided 242 variable characters and 174 informative sites over a total of 399 sites.

Figure 2.3 : One of the MP trees for the *LEAFY* gene with 74 sequences : Detarieae relationships in A and non-Detarieae (clade 1) relationships in B. Statistics of each analysis are given in table 2.5. Dashed lines represent non-supported branches in the strict consensus tree. Numbers above branches represent number of nucleotide substitutions / number of non-synonymous changes in the case of *LEAFY*; Jackknife support above 50% is shown below the branches. AMH: Amherstia; BER: Berlinia; BRA: Brachystegia; BRO: Brownea; CAES: Caesalpinieae; CASS: Cassieae; CER: Cercideae; CRU: Crudia; CYN: Cynometra; DET: Detarieae; HYMN: Hymenostegia; MAC: Macrolobium; MIMO: Mimosoideae; PAPIL: Papilionoideae.

A


26 / 2
Bauhinia bohniana CERC
6 / 0
to clade 1
Anthonotha macrophylla 1
16 / 1
Berlinia confusa 20 BER
Berlinia confusa 21
7 / 0
26 / 13
100
Afzelia bella 13 HYMN
Afzelia bella 3
13 / 4
Brownea coccinea 5 BRO
18 / 7
98
Schotia afra 2
14 / 8
100
Schotia afra 13 CYN
Schotia afra 22
Schotia afra 17
5 / 2
7 / 2
95
Saraca dives 13 HYMN
Saraca dives 10
2 / 0
Humboldtia laurifolia 9
4 / 1
75
Humboldtia laurifolia 7 AMH
Humboldtia laurifolia 2
3 / 1
90
Crudia gabonensis 7 CRU
Crudia gabonensis 11
2 / 0
Hymenostegia ngounyensis 12 HYMN
Scorodophleus zenkerii 9
66
Scorodophleus zenkerii 16 CYN
Brownea multijuga 9
Brownea multijuga 11
4 / 1
74
Brownea multijuga 4
15 / 8
100
55
Brownea multijuga 7 BRO
Brownea grandiceps 2
6 / 4
98
Brownea grandiceps 3
Brownea leucantha 4
55
Brownea leucantha 1
Brachystegia laurentii 4
9 / 3
100
Brachystegia laurentii 2 BRA
Brachystegia mildbraedii 6
Plagiosiphon emarginatus 1
Plagiosiphon emarginatus 12 HYMN
2 / 0
Plagiosiphon emarginatus 3
Cynometra ramiflora 4
7 / 3
90
Cynometra ramiflora 1 CYN
58
Cynometra ramiflora 5
Gossweilerodendron balsamiferum 8
7 / 2
83
Gossweilerodendron balsamiferum 2 CRU
Gossweilerodendron balsamiferum 6
Amherstia nobilis 1
50
Amherstia nobilis 13 AMH
Englerodendron usambarensis 3
Englerodendron usambarensis 2 BER
Englerodendron usambarensis 1
Macrolobium ischnocalyx 4
Macrolobium ischnocalyx 2 MAC
Guibourtia demeusii 4
Guibourtia demeusii 8 CRU
total c. / non-synonynous c.
Jackknife support

B


31 / 19
Dussia tessmanii
17 / 2
32 / 3
Pisum sativum
PAPIL
11 / 0
55
26 / 1
Apios americana 19
100
Apios americana 7
6 / 0
Tamarindus indica 4
9 / 4
97
Tamarindus indica 11
DET
Tamarindus indica 12
3 / 1
18 / 2
Dialium guianensis 5
CASS
7 / 2
Umtizia listeriana 9
4 / 1
Tamarindus indica 7
DET
8 / 3
9 / 1
97
Tamarindus indica 9
54
33 / 10
Caesalpinia decapetala
3 / 2
10 / 2
Schizolobium parahyba 9
Delonix elata 5a
51
CAES
12 / 0
5 / 0
Delonix elata 4
80
Delonix elata 5b
68
Delonix elata 9
6 / 0
63
Brownea coccinea 8
DET
31 / 9
Chamaecrista sp.
CASS
6 / 1
10 /
Senna alata 4
6 / 2
Acacia caven 13
22 / 2
100
Acacia caven 16
MIMO
7 / 2
Entada polyphylla 10
16 / 2
100
Entada polyphylla 12

The analysis recovered 200 000 trees, the maximum retained in memory (L= 696; CI = 0,437; RI=0,794, table 2.5). One of the most parsimonious (MP) trees is shown in figure 2.3 to show branch length (number of total nucleotide substitutions and non-synonymous substitutions) and Jackknife support.

Recovery of a clade including all the Detarieae s.l. (except *Tamarindus* and *Umtiza*) depends on seven nucleotide substitutions, of which none are non-synonymous (Fig. 2.3A). Within this clade, the *Anthonotha* - *Berlinia* clade occurs as sister to other Detarieae, but is not highly supported. The clade containing other Detarieae is recognised because of 18 nucleotide substitutions, of which seven are non-synonymous (Jackknife value of 98%). Within the main Detarieae clade a large polytomy occurs in the strict consensus. Monophyletic groups found in this main Detarieae clade are the *Crudia, Scorodophleus* and *Hymenostegia* group, the *Brownea* group (100% Jackknife); and the *Humboldtia, Plagiosyphon* and *Cynometra* group. The *Guibourtia, Englerodendron* and *Macrolobium* clade is often encountered in individual trees (77%), but not recovered in the strict consensus. As mentioned above, it is noticeable that two *Brownea coccinea* clones do not group together, nor with the other *Brownea* species. One *B. coccinea* clone (clone 5) occurs as sister to the remaining Detarieae clade, while the *B. coccinea* clone 8 is highly supported (63% Jackknife) to form a monophyletic group with *Delonix elata* 9, a Caesalpinieae species (Fig 2.3 B).

In the clade leading to Papilionoideae and Mimosoideae species (clade 1; Fig 2.3 B), resolution is high, but few branches are highly supported. Papilionoideae is resolved as monophyletic (55% Jackknife value), but sampling was extremely small, and *Dussia tessmanii*, a member of the Sophoreae is sister to *Pisum* and *Apios.* This subfamily is not recognised by any non-synonymous character. The Papilionoideae clade is in turn sister to the non-Detarieae Caesalpinioideae and Mimosoideae clade. Mimosoideae *(Acacia* and *Entada* only were sampled) occurs as monophyletic (supported by two non-synonymous characters) and derived from a group comprising of non-Detarieae Caesalpinioideae plus, as mentioned above,

the oddly placed *Brownea coccinea* clone 5 and *Tamarindus* clones. *Tamarindus,* traditionally associated with *Amherstia,* is here nested with a Cassieae species, *Dialium guianensis.*

Table 2.5 Amount of phylogenetic information (number of informative characters) and homoplasy indices (CI and RI) for analyses of different codon positions for the *LEAFY* gene and the *trn*L intron for the Leguminosae.

Analysis	Number of sites	Number of informative characters	Length	CI	RI
total exon	399	174	696	0.437	0.794
1st position only	134	43	114	0.702	0.886
2nd position only	133	25	73	0.877	0.951
1st and 2nd positions only	266	68	197	0.731	0.890
3rd position only	133	107	512	0.395	0.755
without the 32 characters	367	142	442	0.536	0.844
Reweighted	-	174	204,56	0.615	0.893
*trn*L	792	73	155	0.639	0.833
*trn*L + *LEAFY*	1570	327	1072	0.510	0.815

There are a few long branches in the trees. These include terminal branches leading to *Afzelia bella* (15 steps), *Dussia tessmanii* (31 steps), *Pisum sativum* (32 steps), *Caesalpinia decapetala* (33 steps), *Acacia caven* (22 steps) and *Chamaecrista sp.* (31 steps), and the internal branch leading to the main Detarieae clade (clade 2) and the remaining species (18 steps).

Different weighting schemes were analyzed in order to explore analysis stability and homoplasy effect. In phylogenetic analyses using coding sequences, saturation at the third codon position is a matter of concern. This position is often four-fold degenerate and therefore evolves very rapidly because it is not subjected to selective constraints. Saturation signifies that, over a certain evolutionary distance, substitutions occur so frequently that multiple hits (multiple substitutions) lead to reversions in character states. In *LEAFY*, variation at the third codon position is actually very high (89% of variable sites, Table 2.4). Whether different codon positions would be saturated in substitutions was examined by plotting

transitions and transversions in function of total divergence (fig. 2.4). If saturation were present, a plateau effect would be reached in the graph, because, as total divergence increases, there would be no increase in number of transitions or transversions. In contrast, this figure indicates there is no evidence of saturation in either transitions or transversions for any of the three codon positions (Fig 2.4). Despite the apparent absence of saturation, homoplasy was present at the third position because the consistency index (CI) was always less than 0.45 for characters at the third position (table 2.4 and 2.5).

Because most of the variability in the exon is found at the third position, excluding those sites significantly lowers the amount of phylogenetic signal. When we performed a phylogenetic analysis excluding third codon position, only 68 informative characters remained. The strict consensus of the 200, 000 trees was almost completely unresolved grouping mostly clones from single species. This result is intriguing, because mutations at first and second codon positions are more often non-synonymous and are known to evolve at a slower rate than the third codon position, thus they were expected to display a phylogenetic signal at deeper nodes in the trees. Overall, this suggests that these *LEAFY* coding regions have not reached the upper taxonomic limit of their utility.

In addition, six third positions required more than 10 changes and 32 sites needed more than 5 changes when optimized on the most parsimonious trees. When these 32 third positions are excluded, the strict consensus tree is slightly more resolved. An important difference is that *Schotia* is sister to the main Detarieae clade rather than being included within it. A less subjective way of giving lower weight to homoplasious sites is to reweight all characters according to minimal RCI and to reiterate the reweight until the tree length stabilizes. Tree topologies resulting from this analysis are significantly more resolved in the main Detarieae clade, but do not differ much for clade 1. Consequently, we found no justification for excluding all third positions.

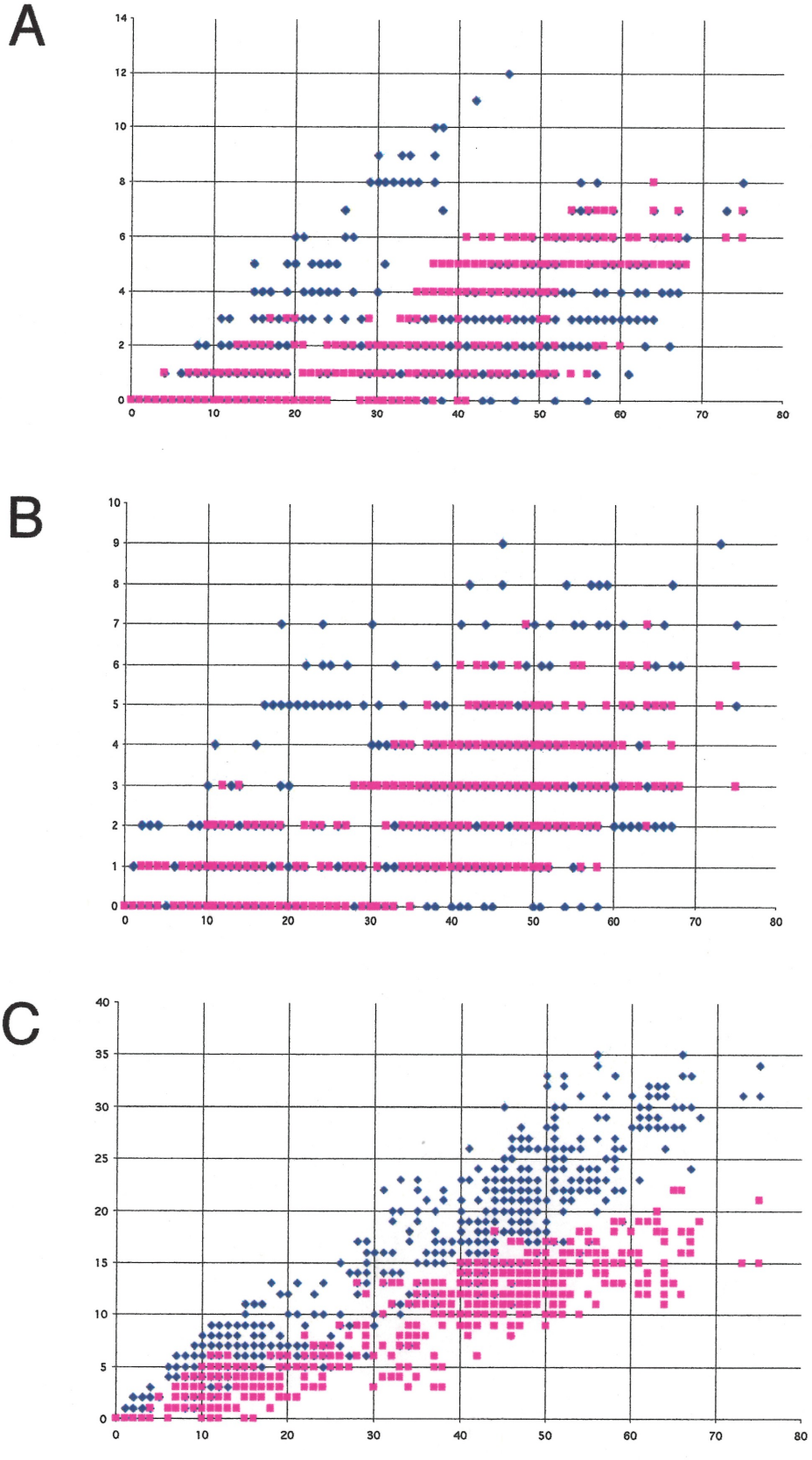
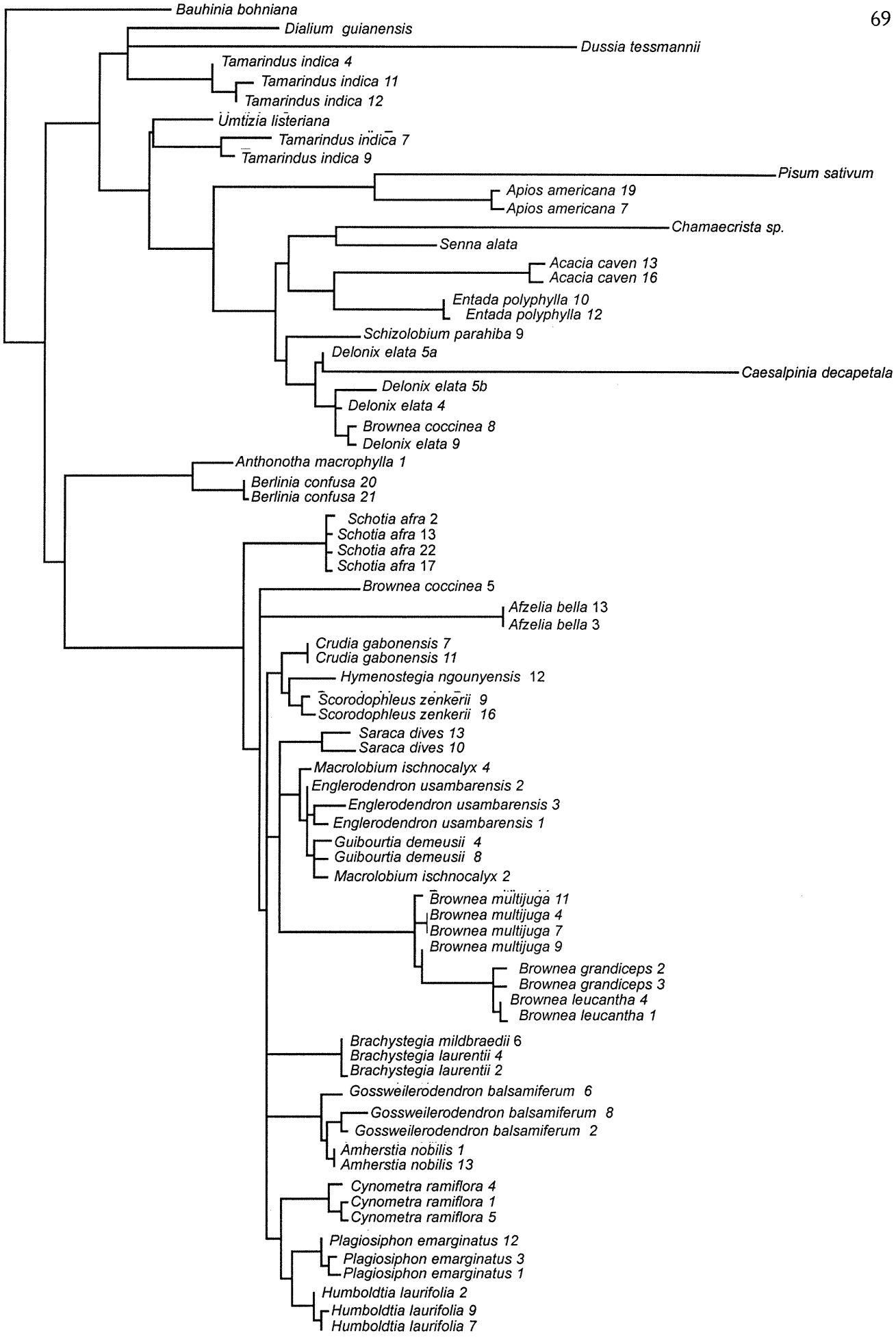


Figure 2.4 : Number of transitions (blue diamonds) and transversions (pink squares) according to position in function of total number of pairwise differences. (A) First position (B) second position (C) third position.

When *Dussia* and *Caesalpinia* are removed from the parsimony analysis because of their long branches, some changes are observed in clade 1 of the resulting trees. *Senna* and *Chamaecrista* are not sister to Mimosoideae but rather paraphyletic and sister to a clade containing most of the non-Detarieae Caesalpinioideae and Mimosoideae. If, in addition, *Pisum* is removed, an interesting change occurs in the main Detarieae clade : *Saraca* is now sister to a clade containing *Guibourtia*, *Englerodendron* and *Macrolobium*. Because of this result, an analysis was done without *Saraca*. It could then be seen that this species had a major impact on tree topology because when excluded, a significant increase in resolution occurs in the main Detarieae clade.

Maximum likelihood (ML) analysis recovered four topologically identical trees (Fig. 2,3C) with the following parameters: -Ln likelihood = 4243,09883; proportion of invariable characters = 0,1; estimated of the gamma shape parameter (alpha) = 0,61258. This value for the shape parameter of the gamma distribution indicates that most of the sites evolve at a slow rate, but that a few sites evolve rapidly. ML and MP trees are similar in several respects, but a few differences can be observed. In the ML tree, *Schotia* clones are sister to the other Detarieae, whereas this position is occupied by *Afzelia* clones in the MP trees. Minor differences between analyses are found in the Detarieae clade. *Dussia tessmanii*, a member of the Sophoreae group (Papilionoideae), is grouped with other Papilioniodeae species in MP trees, but is in a group along with *Dialium guianensis* (Cassieae), and *Tamarindus indica* (Detarieae) species in the ML tree.

Figure 2.5 : ML tree for the same data set as Fig 2.3. Statistics of each analysis are given in table 2.5. Dashed lines represent non-supported branches in the strict consensus tree. Numbers above branches represent number of nucleotide substitutions / number of non-synonymous changes in the case of *LEAFY*; Jackknife support above 50% is shown below the branches.


Bauhinia bohniana
Dialium guianensis
Dussia tessmannii
Tamarindus indica 4
Tamarindus indica 11
Tamarindus indica 12
Umtizia listeriana
Tamarindus indica 7
Tamarindus indica 9
Pisum sativum
Apios americana 19
Apios americana 7
Chamaecrista sp.
Senna alata
Acacia caven 13
Acacia caven 16
Entada polyphylla 10
Entada polyphylla 12
Schizolobium parahiba 9
Delonix elata 5a
Caesalpinia decapetala
Delonix elata 5b
Delonix elata 4
Brownea coccinea 8
Delonix elata 9
Anthonotha macrophylla 1
Berlinia confusa 20
Berlinia confusa 21
Schotia afra 2
Schotia afra 13
Schotia afra 22
Schotia afra 17
Brownea coccinea 5
Afzelia bella 13
Afzelia bella 3
Crudia gabonensis 7
Crudia gabonensis 11
Hymenostegia ngounyensis 12
Scorodophleus zenkerii 9
Scorodophleus zenkerii 16
Saraca dives 13
Saraca dives 10
Macrolobium ischnocalyx 4
Englerodendron usambarensis 2
Englerodendron usambarensis 3
Englerodendron usambarensis 1
Guibourtia demeusii 4
Guibourtia demeusii 8
Macrolobium ischnocalyx 2
Brownea multijuga 11
Brownea multijuga 4
Brownea multijuga 7
Brownea multijuga 9
Brownea grandiceps 2
Brownea grandiceps 3
Brownea leucantha 4
Brownea leucantha 1
Brachystegia mildbraedii 6
Brachystegia laurentii 4
Brachystegia laurentii 2
Gossweilerodendron balsamiferum 6
Gossweilerodendron balsamiferum 8
Gossweilerodendron balsamiferum 2
Amherstia nobilis 1
Amherstia nobilis 13
Cynometra ramiflora 4
Cynometra ramiflora 1
Cynometra ramiflora 5
Plagiosiphon emarginatus 12
Plagiosiphon emarginatus 3
Plagiosiphon emarginatus 1
Humboldtia laurifolia 2
Humboldtia laurifolia 9
Humboldtia laurifolia 7

The position of *Dussia tessmanii* could therefore be considered unstable in the *LEAFY* analyses. The overall similarity of trees using both methods suggests that long branches are not problematic for the cladistic analyses.

The *trn*L intron analysis for the same subset of taxa studied for *LEAFY*, with 281 variable characters and 33 sequences recovered 200,000 trees with a topology highly similar to the one inferred from 239 sequences by (Bruneau et al., 2000; Bruneau et al., 2001). Species relationships as suggested by *LEAFY* differ from the *trn*L analysis in many details, despite being overall alike. Incongruence between parsimony trees from *trn*L and *LEAFY* is revealed by the partition homogeneity (or incongruence length difference, ILD) test (P = 0.01). Important differences between *trn*L and *LEAFY* occur in the main Detarieae clade and some are highly supported. In the *trn*L trees, *Macrolobium* is grouped with *Brownea* rather than with *Guibourtia* and *Englerodendron*. *Guibourtia* is in a clade with *Gossweilerodendron* and *Schotia*, sister to other Detarieae in *trn*L but not in *LEAFY*. In the *trn*L analysis, *Afzelia* is not found sister to the Detarieae as it is with *LEAFY*, but clusters with *Hymenostegia*. In the clade leading to Papilioinoideae and Mimosoideae, there are few differences, but most notable is the absence of *Tamarindus* and *Brownea coccinea*, which occur in the Detariaea clade. A Gene Tree analysis (Page and Charleston, 1997) revealed that 20 *LEAFY* gene duplications would have occurred to reconcile the *trn*L and *LEAFY* trees. Because this high number of duplications is not biologically plausible, we will not discuss this explanation further.

Figure 2.6 : One of the MP tree for the analysis of *trn*L intron sequences. Statistics of each analysis are given in table 2.5. Dashed lines represent non-supported branches in the strict consensus tree. Numbers above branches represent number of nucleotide substitutions / number of non-synonymous changes in the case of *LEAFY*; Jackknife support above 50% is shown below the branches.

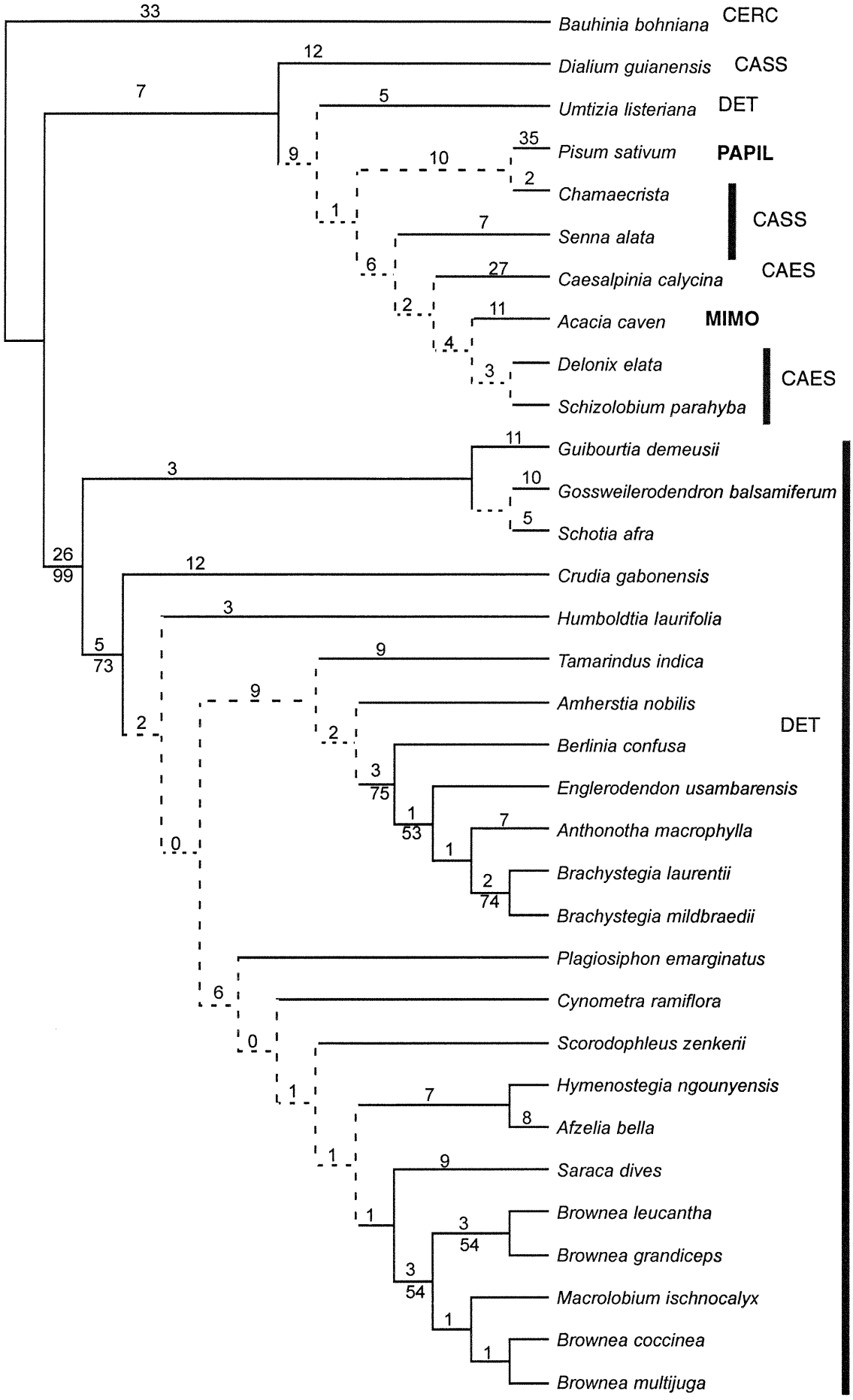
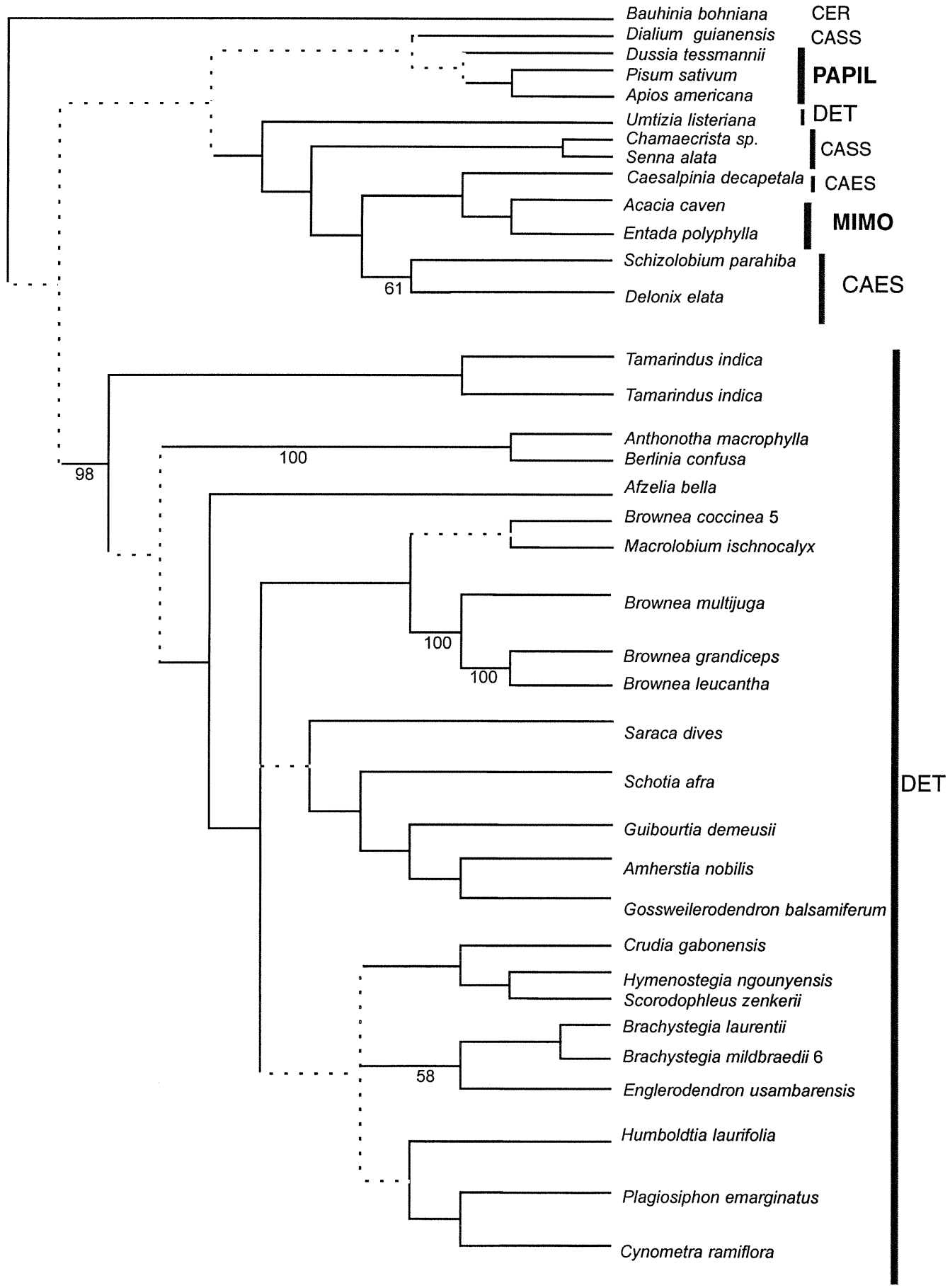

Bauhinia bohniana
CERC
Dialium guianensis
CASS
Umtizia listeriana
DET
Pisum sativum
PAPIL
Chamaecrista
Senna alata
CASS
Caesalpinia calycina
CAES
Acacia caven
MIMO
Delonix elata
Schizolobium parahyba
CAES
Guibourtia demeusii
Gossweilerodendron balsamiferum
Schotia afra
Crudia gabonensis
Humboldtia laurifolia
Tamarindus indica
Amherstia nobilis
Berlinia confusa
Englerodendon usambarensis
Anthonotha macrophylla
Brachystegia laurentii
Brachystegia mildbraedii
Plagiosiphon emarginatus
Cynometra ramiflora
Scorodophleus zenkerii
Hymenostegia ngounyensis
Afzelia bella
Saraca dives
Brownea leucantha
Brownea grandiceps
Macrolobium ischnocalyx
Brownea coccinea
Brownea multijuga
DET

Figure 2.7 : One of the MP tree from the combined analysis of *LEAFY* and *trn*L. Statistics of each analysis are given at table 2.5. Dashed lines represent non-supported branches in the strict consensus tree. Numbers upper branches represent number of nucleotide substitutions / number of non-synonymous changes in the case of *LEAFY*; Jackknife support is shown below branches.


Bauhinia bohniana
Dialium guianensis
Dussia tessmannii
Pisum sativum
Apios americana
Umtizia listeriana
Chamaecrista sp.
Senna alata
Caesalpinia decapetala
Acacia caven
Entada polyphylla
Schizolobium parahiba
Delonix elata
61
CER
CASS
PAPIL
DET
CASS
CAES
MIMO
CAES
Tamarindus indica
Tamarindus indica
Anthonotha macrophylla
Berlinia confusa
100
98
Afzelia bella
Brownea coccinea 5
Macrolobium ischnocalyx
Brownea multijuga
100
Brownea grandiceps
100
Brownea leucantha
Saraca dives
Schotia afra
Guibourtia demeusii
Amherstia nobilis
Gossweilerodendron balsamiferum
Crudia gabonensis
Hymenostegia ngounyensis
Scorodophleus zenkerii
Brachystegia laurentii
Brachystegia mildbraedii 6
58
Englerodendron usambarensis
Humboldtia laurifolia
Plagiosiphon emarginatus
Cynometra ramiflora
DET

Discussion

Phylogenetic utility of the *LEAFY* exons and introns

A gene can be of phylogenetic utility only if the gene tree represents the species tree. Common evolutionary phenomenon in the nuclear genome or in species evolution such as gene duplication, gene conversion or recombination, introgression and hybridization, or character convergence could hamper the capacity of a gene to represent correctly the evolutionary history of species. Further, because different genes evolve at various rates, they can be useful at different taxonomic levels and thus, taxonomic level of utility has to be assessed when trying to use a new DNA sequence in phylogenetic studies. Pertinence of the *LEAFY* gene tree for reconstructing species evolution will first be discussed, including a discussion of the number of copies and of potential paralogy problems. Variation in the coding sequence of *LEAFY* then will be explored in relation to taxonomic level. Finally, congruence of phylogenetic history among independent data sets is an important criterion for robustness of phylogenetic hypothesis (Friedlander et al., 1996). Therefore, the topological congruence of *LEAFY* with the *trn*L chloroplast intron analysis will be assessed (Bruneau et al., 2001), and explanations of incongruences, where encountered, will be explored.

Number of copies

The confirmation that *LEAFY* is a single copy gene in the Caesalpinioideae is crucial for the proper interpretation of the topology of the phylogenetic tree. Multiple copies could be subject to recombination or gene conversion, or paralogy-orthology confusion (Wendel and Doyle, 1998). However, if the gene occurs in a single copy, such mechanisms could not explain eventual incongruence between *LEAFY* and other phylogenetic markers.

Results showing that *LEAFY* is a single copy gene are expected, considering that *LEAFY* has been found as a single copy gene in all diploid angiosperms studied to date (Frohlich and Parker, 2000; Theissen, 2000), including *UNIFOLIATA,* the *LEAFY* homolog in *Pisum sativum* (Hofer et al., 1997). In

gymnosperms, two copies of *LEAFY* are found and the angiosperm *LEAFY* is thought to originate from one of these copies, while the other is lost (Frohlich and Parker, 2000). However, considering the paleopolyploid nature of a large proportion of angiosperms (Masterson, 1994) and the fact that most nuclear genes are members of multigene families (Vision et al., 2000), duplication being a common process, establishing the number of copies of *LEAFY* in the Caesalpinioideae remains an important consideration.

What do our results suggest regarding the number of copies of *LEAFY* in the Leguminosae? Tree topology and divergence between clones are not conclusive in determining copy number of *LEAFY* in the Caesalpinioideae. When clones of a same species form a monophyletic group (or paraphyletic because of lack of resolution), divergence is very low (less than 1%), and could well be explained by divergent alleles at a single locus, rather than by the presence of multiple loci. However, the presence of clones of *Brownea coccinea* in two distant clades is suggestive of the presence of two copies of *LEAFY* in this species, unless *B. coccinea* is an allotetraploid formed after hybridization between a Detarieae species and a Caesalpinieae species, close to *Delonix elata*. Although chromosomes counts are not known for *Brownea coccinea*, polyploidy is possible but hybridisation between very divergent genera seems unlikely. The fact that *Tamarindus* clones form two distinct monophyletic clades not very distant one from the other suggests that there are two different loci of *LEAFY* in this species. Similarly, polyploidy of this widely cultivated species could explain this tree topology, where both loci would have to have diverged just enough to form distinct clades. However, polyploidy in *Tamarindus* does not in itself provide an explanation as to why this Detarieae species does not group with other Detarieae. Thus, although results are still unclear about the copy number of *LEAFY* in the Caesalpinioideae, they demonstrate that this gene is most probably in single copy, except possibly in *Brownea coccinea* and *Tamarindus indica*.

Intron sequence utility:

The type of mutations in the second intron of *LEAFY* is similar to the mutational patterns found in other non-coding sequences, such as in the chloroplast intron (Kelchner, 2000). In contrast to the chloroplastic *trn*L intron analysis (Bruneau et al., 2001), the second *LEAFY* intron presents an extremely high level of variation at the family level (Fig. 2.2), making homology assessment for alignment not reliable between tribes and subfamilies of the Leguminosae. As noted in other studies (Small et al., 1998), alignability is an important issue when analyzing and choosing between coding and noncoding regions for phylogenetic studies. Because noncoding regions accumulate indels of all lengths at high frequency compared to coding regions, which are constrained to maintain the reading frame, sequence alignment is more problematic in introns. The level of variation in the intron at the family level is comparable to that observed at the third codon position, where almost all sites are variable. However, unlike for the intron sequences, homology of the third codon is easier to assess because of positional homology. In numerous other studies of nuclear genes (e. g. Mason-Gamer et al., 1998; Gaut et al., 1999; Evans et al., 2000; Koch et al., 2000; Oxelman and Bremer, 2000), introns were declared not useful because the degree of variation was too high for the taxonomic level studied.

At the species level however, the second intron of *LEAFY* may not be as variable as expected. Within the Leguminosae, the level of variation offered by the second intron of *LEAFY* may prove useful for phylogenetic analyses at a very narrow range of taxonomic levels, at the tribal level or slightly below. Above this level, the intron is too variable in length and in sequence to be aligned, while at the species level it may not be variable enough. Between the two *Brachystegia* species studied, for instance, only seven bases differ. The taxonomic level at which the utility of the intron appears optimal is between genera of the tribe Detarieae. It is only between distantly related genera that an overlap in phylogenetic utility of the exon and of the intron sequences exist (see below for exon utility). Nonetheless, this intron represents an interesting phylogenetic marker, because of its characteristic of being a structurally conserved, noncoding

and single copy region from the nuclear genome of which few are available for phylogenetic analysis. Few other noncoding sequences of this type have been used with success in phylogenetic studies between species or closely related genera, including the *pistillata* intron in *Sphaerocardamum* (Bailey and Doyle, 1999), the *H3D* intron in *Glycine* (Doyle et al., 1996b) or *ncpGS* in *Oxalis* (Emshwiller and Doyle, 1999).

Discovery of a new intron in *Brownea* species

The presence of a new intron suggests that intron gain may serve as a taxonomic marker in the genus *Brownea*. Caution should be exerted, however, when using presence or absence of an intron as a taxonomic character, because no clear and unambiguous mechanism has been proposed for intron gain or loss. An alternative hypothesis to this particular gain, to explain presence of an additional intron in some *Brownea* species, would be that this intron was present in the first organism with a *LEAFY*-like gene, but was lost in all lineages investigated to date except *Brownea*, as the intron early theory would suggest (Gilbert, 1978; Darnell and Doolittle, 1986). Intron losses are explained by the reverse transcription of a cellular mRNA of the gene into the genome, with gene replacement by homologous recombination events (Baltimore, 1985). It would be of interest to determine whether the new intron we discovered is located between protein domains of LEAFY. If so, this would give some credit to the exon origin of intron theory. However, in the case of this specific intron, recent insertion appears more likely because innumerable independent losses would be needed to account for the phylogenetic distribution of the intron. Supporters of the insertional theory of introns (Cavalier-Smith, 1985; Patthy, 1985; Rogers, 1985) suggest that intron insertion can account most likely for position and phylogenetic distribution of the intron in a lot of cases. Very few intron gains are reported. They are reported in the catalase (*CAT*) of *Oryza* (Frugoli et al., 1998; Iwamoto et al., 1998), in triose phosphate-isomerase (*Tpl*) of insects (Logsdon et al., 1995), *rbcS* of Solanaceae, and actins and serine protease (Patthy, 1985; Rogers, 1985). However, theories about intron appearance are still a matter of debate.

Intron gains are rarely encountered, and are even more rarely used from a taxonomic perspective (Iwamoto et al., 1998; Venkatesh et al., 1999). When using intron presence or absence as a taxonomic character, we should be alert to the possibility of multiple gains or losses. Indeed, parallel intron losses have been demonstrated empirically in a number of taxa. In the Brassicaceae, the fourth intron of *ADH* was lost independantly twice (Charlesworth et al., 1998; Koch et al., 2000), and all of the introns were lost independently in *ADH2* of *Arabis procurrens, A. blepharophylla, A. hirsuta* (Koch et al., 2000), and *ADH*3 of *Leavenworthia* (Charlesworth et al., 1998).

Dibb and Newman (1989) suggested that new introns, like those found in tubulin genes, insert in a consensus MAGR site called proto-splice site. In our study, the new 490 aligned bp intron has inserted in a site that has a AAKG sequence, compatible - although not exactly the same - with a proto-splice site. It has been suggested that an intron can be gained when a transposable element is inserted in the gene and then correctly spliced (Giroux et al., 1994), or by the duplication of a portion of the coding sequence containing a proto-splice site (Venkatesh et al., 1999). Because the sequence of this intron is not similar to any other known sequence, nor to any mobile element or a portion of the *LEAFY* exons, no clue exists as to its origin.

This new intron subdividing the second exon of the *LEAFY* gene could represent a valid polarized taxonomic character because it occurs in closely related *Brownea* species. If we consider the acquisition of a new intron as a polarized character, the new intron found in the *Brownea* group suggests *B. grandiceps*, *B. multijuga*, and *B. leucantha,* which possess the intron, to be more closely related to each other than to *Brownea coccinea,* which lacks the intron. This relationship is not congruent with that suggested by the *trn*L intron sequence evolution (Bruneau et al., 2000). The presence of this additional intron could help in delimitating groups within the genus *Brownea*, or in the assignment to a particular group of an unidentified individual. Indeed, in this genus, many taxonomic changes have been proposed, but relationships remain under study.

For example, *Brownea capitella* is considered a subspecies of *Brownea coccinea* (Velazquez, 1992). This hypothesis could be verified by the absence of the additional intron in *LEAFY* gene. Also, although numerous characters separate *Browneopsis* from *Brownea* (Klitgaard, 1991), at least two species have been transferred from one genus to the other in the past. One of these is *Browneopsis disepala*, which was transferred from *Brownea* to *Browneopis* by Klitgaard (1991). Presence or absence of the intron could help classify unambiguously taxa in this case.

The lower level of variation observed in this intron compared to other intron sequences in our study, may be an indication that this intron is of more recent origin and have been inserted in a single individual, ancestral to *Brownea multijuga, B. grandiceps and B. leucantha*. This is in contrast, for example, of the second *LEAFY* intron, which may have been polymorphic at the moment of speciation and thus, may have conserved a higher variation in sequence via allelic recombination. Further, we hypothesize that sequence variation in this new intron could provide phylogenetic information within this genus.

Sequence variation in the exons

In the case of *LEAFY*, the complete amino acid sequences proved to be alignable and phylogenetically useful among all spermatophytes (Frohlich and Meyerowitz, 1997; Frohlich and Parker, 2000); that is to say, relationships were reliable and well supported. This first study of *LEAFY* gene evolution clearly showed the third exon as being the most conserved part of the entire gene. The ideal amount of variation a gene should possess to be useful - while remaining reliable - is difficult to determine. Empirically, most of the phylogenetic studies using nuclear coding sequences work with data sets showing up to 50% variable sites (Fig 2.5). *LEAFY* shows a similar proportion of variation (56%) compared to other nuclear genes used at the family level (Fig. 2.5), slightly higher than the variability exhibited by the *GBSSI* gene in the Poaceae (51%) (Mason-Gamer et al., 1998) and the vicilin genes in the Sterculeaceae (42%) (Whitlock and Baum, 1999). It also experiences similar problems such as a high homoplasy content

revealed by a low consistency index (Mason-Gamer et al., 1998; Mathews and Donoghue, 1999), and lack of support.

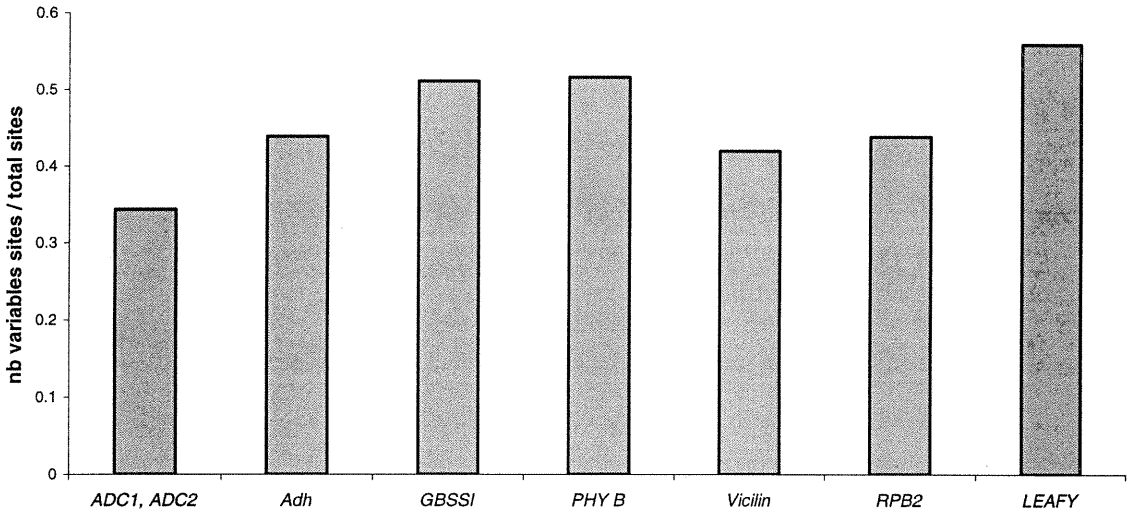


Figure 2.8: Proportion of variable sites in some data sets of nuclear genes used at the family level. Statistics come from Brassicaceae (Galloway et al., 1998; Koch et al., 2000), Poaceae (Mason-Gamer et al., 1998; Mathews et al., 2000), Sterculiaceae, (Whitlock and Baum, 1999), Gentianales (Oxelman and Bremer, 2000) and Leguminosae (this study).

The high level of variation of *LEAFY* in the Leguminosae provides phylogenetic signal at and below the family level. Although the Leguminosae database of *LEAFY* sequences shows more than 50% of variable sites, the phylogenetic utility of *LEAFY* nucleotides sequence at the family level is still reliable because saturations of substitutions was not detected. Between distantly related genera, as those within the clade 1 for exemple (Fig. 2.2B), sequence variation and resolution are high, and *LEAFY* is thus very effective at that level. However, within tribe Detarieae, the lack of resolution in individual trees (Fig 2.2A) reveals a probable limit to phylogenetic utility of *LEAFY* just above the generic level in this group. Our data from the third exon suggest that amino acid sequence is not variable enough across the family, but that nucleotide sequences from coding regions could be used for a broad range of different taxonomic levels, with a high performance between distantly genera or at a taxonomic level that is slightly

broader than the Detarieae s.l. There is a possibility for this region to be used for phylogeny reconstruction among related families although in this case, third positions may be subject to extensive saturation.

As expected for a gene required for flowering in many angiosperms, *LEAFY* sequence evolution is conservative, and non-synonymous substitutions are less common than synonymous ones in the Leguminosae. Because the ratio is calculated across the entire sequence, this simple statistic may lack power to detect local adaptative events. From the alignment, it can be observed that some sites and regions are more conserved that others, but because functional domains of the LEAFY protein have not been characterised, it is impossible to link conservation within the *LEAFY* sequence with any particular function of that protein. In the MADS-box genes, another family of transcription factors involved in development, the most variable regions may serve in transcriptional activation, and the most conserved ones are involved in DNA binding and dimerization (Riechmann and Meyerowitz, 1997). However because the MADS-box is a large multigene family of about fifty members in *Arabidopsis* and is characterised by functional redundancy, and because their proteins are often acting as dimers, comparison with *LEAFY* may not hold completely.

Congruence with another DNA region: the chloroplast *trn*L intron

Concerning subfamilial and tribal relationships, the *LEAFY* tree is congruent overall with analyses from the chloroplast intron *trn*L. However, in the Detarieae s.l., the *LEAFY* phylogeny shows striking incongruence with the *trn*L phylogeny. Many groupings suggested by *LEAFY* do not reflect either results from the chloroplast analyses (Bruneau et al., 2000; Doyle et al., 2000; Bruneau et al., 2001) or taxonomic classifications based on morphological characters (Polhill et al., 1981; Polhill, 1994). Conflicts in species trees implied by different molecular data can be of two origins. They may be methodological artefacts or they can reflect a highly complex phylogenetic history of the genes.

i) Methods of phylogenetic analyses

One possibility to explain the marked incongruence between *LEAFY* and *trn*L reconstructions is that the model of sequence evolution subtending the phylogenetic reconstruction method used dramatically differs from the evolution of the region studied, a problem that may be exacerbated by suboptimal sampling of species in the *LEAFY* Leguminosae data set. Examples of sequence evolution characteristics that could drive the analysis to not recover the true tree are GC content and transition transversion ratios, deviation from stationarity of nucleotide frequency (Mason-Gamer et al., 1998), among lineage rate heterogeneity or among site rate heterogeneity (Yang, 1996; Yang and Kumar, 1996). The *LEAFY* data for the Leguminosae possesses some of those characteristics : among site and among lineage rate heterogeneity. However, the overall similarity of MP trees with the ML trees, where all parameters are estimated from the data, suggests that these factors are not the cause of incongruence. Other nuclear genes show this kind of incongruence with chloroplast data. Phylogenies resulting from *RPB2-d* sequences in the Gentianales were not congruent with *ndh*F or *rbc*L chloroplast sequences (Oxelman and Bremer, 2000). In the Rosaceae, two copies of *GBSSI* do not result in the same species phylogeny (Evans et al., 2000) and in the Triticaceae, the *GBSSI* data are incongruent with chloroplast regions (Mason-Gamer et al., 1998). In both *GBSSI* studies, heterogeneity of the substitution pattern among genes, among sites or among lineages were the most likely causes of incongruence. *LEAFY* and the *trn*L intron are expected to be subject to different substitution patterns since they are from different genomes and since one is a coding sequence and the other is not. Moreover, the sampling of *LEAFY* sequences in the Leguminosae is heterogeneous, concentrating on the tribe Detarieae, but including also highly derived species from the other subfamilies. This sampling could be problematic in phylogenetic reconstructions if the evolutionary patterns of the sequences are not easily analysed. Important sampling effect on tree topology is suspected based on the observation that the removal of some particular sequences from the analysis changes relationships in the remaining taxa (see results).

ii) Undetected multiple copies

Another methodological cause of incongruence between *LEAFY* and a chloroplast sequence may be an incomplete sampling of paralogs and orthologs in a multiple copy gene data set. The possibility that *LEAFY* occurs in multiple copies can not be excluded entirely based on our study, but the presence of more than one copy in every Leguminosae species is not suggested by our data. However, if multiple copies are present in, for example, *Brownea coccinea* and *Tamarindus indica*, these copies seem not to be pseudogenes, although expression data is lacking to confirm this. A possibility exists that there was once more than one copy of *LEAFY* in the Leguminosae, but that different copies became pseudogenes during the evolution of the family. This would result in an apparent unique copy of *LEAFY*, and would create incongruence with chloroplast data. A thorough sampling in the most basal lineages, such as the Cercideae or the Detarieae, could provide some clues to this important question. Detection of pseudogenes, with genomic libraries, or Southern blots may also be helpful.

iii) Introgression or hybridisation, and lineage sorting

Biological processes such as introgression or lineage sorting can often explain incongruence between gene trees (Wendel and Doyle, 1998). In that case, the observed incongruence reflects the hybrid origin of the introgressed species or the ancestral polymorphism randomly sorted in the case of lineage sorting. Both processes occur at low taxonomic levels and would result in the same pattern of incongruence (Wendel and Doyle, 1998). They are thus less probable explanations for the incongruence between the *LEAFY* and *trn*L phylogenetic analyses of the Caesalpinioideae. Indeed, members of this subfamily have been morphologically distinct for a very long time (Herendeen et al., 1992; Herendeen and Jacobs, 2000). Apart from the taxonomic level at which incongruence is found, biogeography is another factor that weakens the introgression and lineage sorting explanations. The most improbable clades suggested by *LEAFY*, the groupings of *Gossweilerodendron* with *Amherstia* or of *Englerodendron* with *Macrolobium* for example, involve species that occur on

different continents: Africa (*Gossweilerodendron* and *Englerodendron*), America (*Macrolobium*) or Asia (*Amherstia*).

However, hybridisation and lineage sorting have been postulated in a case of incongruence between trees from chloroplast and nuclear sequences or between different nuclear DNA sequences (Mason-Gamer et al., 1998) of the tribe Triticaceae (Gramineae) (Kellogg et al., 1996). Triticaceae and Caesalpinioideae are both large tribes from among the largest angiosperms families. Kellogg et al. (1996) suggested there must have been an ancestral polymorphism and a burst of diversification for lineage sorting to be a possible explanation of present incongruence between chloroplast DNA and nuclear DNA. If hybridisation occured, it must have been following a rapid diversification and have ceased soon after. Such a rapid diversification in the Detarieae can be inferred from a polytomy in both *trn*L and *LEAFY* trees at the Detarieae node. Thus, introgression or lineage sorting - although not expected among divergent genera of the Detarieae s. l. - cannot be excluded completely, and is still a valid biological explanation of incongruence in that tribe.

iv) Convergence

One could imagine that the LEAFY protein, because of a partial redundancy in its function with other transcription factors like AP1 (Shannon and Meeks-Wagner, 1993), is under less evolutionary constraint than other important genes such as *RPB2*. Some of the most variable floral characters of the Detarieae, such as petals formation, are under the possible control of *LEAFY*. We wonder whether those characters could be associated with pollination syndromes, that are known to show reversions or shifts in their evolution (eg. Bruneau, 1997; Pennington et al., 2000), and that those differences could be reflected in the *LEAFY* sequences. Furthermore, some floral characters may be associated with particular pollination system, thus showing reversions or shifts in evolution. In *Arabidopsis* and *Pisum*, *LEAFY* (*UNIFOLIATA* in *Pisum*) has a major effect on petal and stamen development (Schultz and Haughn, 1991; Singer et al., 1999); for example, in *lfy* or *uni* mutants, sepal or carpel intermediates are sometimes formed, but never

petals or stamens. In the basal Leguminosae and especially in Detarieae, the second and third (petals and stamens) whorls are extremely variable, showing losses and suppression (Tucker, 1987; Tucker, 2000b; Tucker, 2000a; Tucker, 2001), or proliferation of stamens (Tucker, 1987). Leaf morphology, which is also variable among the Caesalpinioideae in terms of dissection, is another process regulated by *LEAFY* expression and interactions with other genes in the legumes. Alteration in the *UNIFOLIATA*, *AFFILA* or *TENDRILLESS* genes in pea can transform leaves into forms that occur in other Leguminosae such as pinnate or trifoliolate leaves (Hofer and Ellis, 1998).

Could the *LEAFY* sequence evolution differ significantly from neutrality, being correlated with the morphological evolution of leaves and flowers rather than reflecting speciation and evolution of species? Such hypotheses are difficult to test. Searches for a link between sequence evolution and floral morphology evolution have been undertaken for other floral homeotic loci like the MADS-box genes in angiosperms (Kramer et al., 1998; Lawton-Rauh et al., 1999) or the cycloidea genes involved in flower symmetry (Citerne et al., 2000). Conclusions were that changes in homeotic gene sequences seemed to play a secondary role in morphological diversification, compared with changes in gene expression (Kellogg, 1996; Lawton-Rauh et al., 1999). This conclusion is supported by the finding that the expression of a gymnosperm sequence can functionally replace the original LEAFY protein in an *Arabidopsis* plant (Theissen, 2000). Because the biochemical function of the partial region sequenced in *LEAFY* and expression patterns of *LEAFY* in the Caesalpinioideae are still unknown, any correlation is even more difficult to assess and remains highly speculative. Thus, the hypothesis of convergence of LEAFY sequences to explain incongruence between *LEAFY* and other markers is not that likely.

Distinction between causes of incongruence between *LEAFY* and *trn*L will come from the comparison with other nuclear markers in the Leguminosae. If the convergent evolution of *LEAFY* sequences or disagreement between models of analysis and sequence evolution were the only reason of incongruence,

incongruence should not occur when comparing trees from other nuclear data and chloroplast DNA. But if introgression or hybridisation happened, incongruence is expected from most of the comparisons between nuclear and chloroplast trees.

Conclusion

The evolution of coding sequences of nuclear genes, added to the specific complexity of evolution of the whole nuclear genome, make it difficult to use this genome as a single source of phylogenetic characters. Nonetheless, in combination with other markers, *LEAFY* can increase resolution in a phylogenetic analysis. *LEAFY* is shown to be in a single copy in the Leguminosae, except for two species, for which polyploidy is suspected. Exon sequences are also shown to be useful at and below the familial taxonomic level. However, convincing explanations of the incongruence between *LEAFY* and *trn*L sequences in the Detarieae have not been found. Biological explanations such as hybridisation, introgression or functional convergence seem unlikely, but cannot be excluded. Clearly, more information on the functional properties of the protein are needed for testing the convergence hypothesis. Evolution of *LEAFY* sequences may be incompatible with models underlying phylogenetic analysis, although this remains difficult to test. This incompatibility could originate from an insufficient sampling. Utilisation of more nuclear loci could help discriminate between different explanations, as well as increase information on the Leguminosae evolution. LEAFY well illustrates the difficulties of undertaking molecular systematic studies using nuclear sequences. Efforts are rewarded by fascinating discoveries. The evolution of the basal members of the Leguminosae is still unclear after the *LEAFY* phylogenetic analysis. However, the new intron discovered is a potentially excellent character to group some species of the genus *Brownea,* and need further exploration.

Conclusion générale

Les résultats de cette étude ont démontré que la séquence de *LEAFY* pourrait être utile pour la taxonomie des légumineuses. La variabilité des séquences des exons est suffisante pour procurer des reconstructions phylogénétiques bien résolues et supportées au niveau taxonomique de la famille. L'utilité au niveau de la famille n'avait encore jamais été testée pour ce gène, son efficacité n'était connue qu'entre les plantes vertes. De plus, la variabilité de *LEAFY* au niveau de la famille, qui est semblable à celle des autres gènes utilisés aussi à l'intérieur d'une famille, confirme son utilité à cette échelle taxonomique. Les reconstructions sont peu résolues entre espèces de la tribu des Detarieae, mais ce manque de résolution est aussi obtenu lors de l'analyse avec d'autres séquences. Ce gène possède des caractéristiques communes à beaucoup de régions codantes, notamment un taux d'évolution extrêmement rapide à la troisième base du codon. Le second intron de *LEAFY* est suffisamment variable pour fournir de l'information entre genres ; entre espèce, la variation dans l'intron est très faible.

Plusieurs aspects restent encore flous. L'incongruence marquée entre les analyses de *LEAFY* et de *trn*L pour les Detarieae n'est pas pleinement expliquée. L'incongruence entre les données chloroplastiques et nucléaires est courante, *GBSSI* et *RPB2* en sont de bons exemples. Dans le cas présent, par élimination, l'incompatibilité entre l'échantillonnage réduit, l'évolution du gène et le modèle d'évolution implicite à la méthode d'analyse semble la cause la plus plausible pour expliquer l'incongruence. Les patrons de substitutions des différents gènes pourraient, s'ils sont suffisamment différents, mener à des reconstructions phylogénétiques différentes même si les deux régions d'ADN ont subi les mêmes patrons d'embranchement. Ces différences se retrouvent peut-être entre *LEAFY* et *trn*L et il est extrêmement difficile d'échapper aux effets adverses de cette situation sans un échantillonnage intensif. Cette explication est avancée par les auteurs de plusieurs études pour expliquer l'incongruence. Nos données n'incluent qu'une infime proportion des espèces de légumineuses existantes, les

analyses pourraient donc être sensibles à ce mécanisme menant à de l'incongruence.

Bien sur, l'échantillonnage devra être augmenté considérablement. Quelles seront les conséquences sur la classification des légumineuses ? On s'attend à ce que les sous-familles des Caesalpinioideae et de Mimosoideae ne soient plus reconnues. Les Caesalpinioideae sont paraphylétiques dans toutes les analyses publiées jusqu'ici et le monophylétisme des Mimosoideae est encore à démontrer. Qu'en sera-t-il pour les tribus et groupes des Caesalpinioideae ? Les plus récentes analyses phylogénétiques moléculaires ne sont déjà pas en accord avec les classifications traditionnelles, mais les analyses de LEAFY suggèrent une toute autre histoire. Comment réconcilier ces différentes hypothèses dans un contexte de classification ?

Comme pour les autres espèces échantillonnées avant notre étude, *LEAFY* serait en copie unique dans les légumineuses sauf pour deux exceptions. Les gènes en copie unique sont extrêmement rares dans les génomes végétaux. Parmi les gènes utilisés en systématique moléculaire des plantes, décrit à la première section, aucun n'est en copie unique dans tous les diploïdes étudiés. Certains gènes tels que *RPB2*, viciline, *GBSSI* sont en copie unique dans la plupart des espèces, mais avec des exceptions pour certains groupes taxonomiques où plus d'une copie est trouvée. D'autres gènes comme *GAPA*, *ADC* ou les phytochromes sont membres de familles multigéniques, où les différents membres évoluent indépendamment les uns des autres. Ils ne subiraient pas de conversion ou de recombinaison, et leur comportement en analyse phylogénétique serait comparable à celui d'un gène en copie unique. Pour cet aspect, le gène *LEAFY* semble unique jusqu'ici.

La fonction de LEAFY nous amène aussi à nous poser des questions d'un ordre complètement différent. Chez *Pisum*, ce facteur de transcription promeut la détermination des méristèmes d'inflorescence en permettant la transition au méristème floral et il maintient l'indétermination permettant la formation de feuilles

composées dans les méristèmes foliaires. Compte tenu que la diversité de la forme des feuilles et des fleurs, et du moment de floraison est prodigieuse, surtout chez les légumineuses les plus anciennes, il sera passionnant de connaître le rôle que LEAFY joue dans le développement de ces diverses formes. Il faudra connaître les éléments régulateurs, les lieux d'expression, et les interacteurs de la protéine afin de comprendre le rôle de LEAFY dans le développement d'espèces proches parentes, mais morphologiquement différentes. Par exemple, plusieurs espèces de la sous-famille des Cercideae ont des feuilles entières alors que leurs proches parents, sans doute des membres des Detarieae, ont plus souvent des feuilles composées. D'ambitieux projets de recherche ont déjà été proposés afin de connaître l'évolution du développement des feuilles des légumineuses.

Un autre résultat important de la présente recherche laisse entrevoir de nombreuses questions. La séquence du nouvel intron découvert dans le second exon de *LEAFY* chez quelques espèces de *Brownea* ne donnait aucun indice sur le mécanisme d'insertion qui l'a amené à cet endroit. Puisque aucun des mécanismes connus jusqu'ici ne semblait expliquer la présence du nouvel intron, plus d'ombres que de lumières reste encore sur ce sujet. Peut-être la séquence de l'intron se retrouve-t-elle ailleurs dans le génome des espèces porteuses de l'intron ? Les expériences pour le déterminer pourraient être simples et très informatives, et procureraient de l'information nouvelle sur l'évolution des génomes dans les Caesalpinioideae. D'autre part, l'intérêt taxonomique de la présence de l'intron est immense. Parce que l'insertion d'un nouvel intron est un événement rare dans l'évolution des génomes il pourrait représenter un caractère exempt d'homoplasie. Par exemple, si l'intron est retrouvé dans des espèces ou des genres apparentés à *Brownea,* comme le laisse entrevoir des expériences préliminaires, faudra-t-il réviser la taxonomie de ces taxons ? Quel devrait être le poids de ce caractère comparativement aux autres ? Les résultats de ce mémoire de maîtrise prouvent que le génome nucléaire a un incroyable potentiel en analyses phylogénétiques : il procure un grande quantité de caractères, variés et pouvant être utiles à plusieurs niveaux taxonomiques.

Bibliographie

Adachi, J., K. Kosuge, T. Denda et K. Watanabe. 1995. Phylogenetic relationships of the *Berberidaceae* based on partial sequences of the gapA gene. *Plant Systematics and Evolution* Supplement 9, 351-353

Alvarez-Buylla, E. R., S. Pelaz, S. J. Liljegren, S. E. Gold, C. Burgeff, G. S. Ditta, L. Ribas de Pouplana, L. Martinez-Castilla et M. F. Yanofsky. 2000. An ancestral MADS-box gene duplication occured before the divergence of plants an animals. *Proceedings of the National Academy of Sciences of the United States of America* 97 (10), 5328-5333

Bailey, C. D. et J. J. Doyle. 1999. Potential phylogenetic utility of the low-copy nuclear gene *pistillata* in dicotyledonous plants: comparaison to nrDNA ITS an *trn*L intron in *Sphaerocardamum* and other Brassicaceae. *Molecular Phylogenetics and Evolution* 13 (1), 20-30

Baldwin, B. G., M. J. Sanderson, J. M. Porter, M. F. Wojciechowski, C. S. Campbell et M. J. Donoghue. 1995. The ITS region of nuclear ribosomal DNA: a valuable source of evidence on angiosperm phylogeny. *Annals of the Missouri Botanical Garden* 82, 247-277

Baltimore, D. 1985. Retroviruses and retrotransposons: the role of reverse transcription in shaping the eucaryotic genome. *Cell* 40, 481-482

Bennett, M. D. et I. J. Leitch. 1995. Nuclear DNA amount in angiosperms. *Annals of Botany* 76, 113-176

Berry, A. et A. Barbadilla. 2000. Gene conversion is a major determinant of genetic diversity at the DNA level. *Dans* R. S. Singh et C. B. Krimbas (eds), Evolutionary Genetics. Cambridge University Press, Cambridge, pp. 102-123.

Blazquez, M. A. et D. Weigel. 2000. Integration of floral inductive signals in *Arabidopsis*. *Nature* 404, 889-892

Bremer, K., M. W. Chase, P. F. Stevens, A. A. Andberg, B. A., B. Bremer, B. G. Briggs, P. K. Endress, M. F. Fay, P. Goldblatt, M. H. G. Gustafsson, S. B. Hoot, W. S. Judd, M. Kallersjo, E. A. Kellogg, K. A. Kron, D. H. Les, C. M. Morton, D. L. Nickrent, R. G. Olmstead, R. A. Price, C. J. Quinn, J. E. Rodman, P. J. Rudall et V. Savolainen. 1998. An ordinal classification

for the families of flowering plants. *Annals of the Missouri Botanical Garden* 85 (4), 531-553

Brown, J. W. S. 1996. *Arabidopsis* intron mutations and pre-mRNA splicing. *The Plant Journal* 10 (5), 771-780

Brown, J. W. S., P. Smith et C. G. Simpson. 1996. *Arabidopsis* consensus intron sequences. *Plant Molecular Biology* 32, 531-535

Bruneau, A. 1997. Evolution and homology of bird pollination systems in *Erythrina* (Leguminosae: Phaseoleae). *American Journal of Botany* 84, 54-71

Bruneau, A., F. J. Breteler, J. J. Wieringa, G. Y. F. Gervais et F. Forest. 2000. Phylogenetic relationships in tribes Macrolobieae and Detarieae as inferred from chloroplast *trn*L intron sequences. *Dans* P. S. Herendeen et A. Bruneau (eds), Advances in Legume Systematics, part 9. Royal Botanic Gardens, Kew, pp. 121-150.

Bruneau, A., F. Forest, P. S. Herendeen, B. B. Klitgaard et G. P. Lewis. 2001. Phylogenetic relationships in the Caesalpinioideae (Leguminosae) as inferred from the chloroplast *trn*L intron sequences. *Systematic Botany* 26 (3), 487-514

Buchanan, B. B., W. Guissem et R. L. Jones. 2000. *Biochemistry and the molecular biology of plants*. The American Society of Plant Physiologists, Rockville, Maryland. 1367 p.

Cavalier-Smith, T. 1985. *Nature* 315, 283-284

Center for Biological Sequence Analysis. 2001. *Net Plant Gene*. [En ligne]. http://www.cbs.dtu.dk/services/NetPGene/ (15 février 2001).

Charlesworth, D., F.-L. Liu et L. Zhang. 1998. The evolution of the alcohol dehydrogenase gene family by loss of introns in plants of the genus *Leavenworthia* (Brassicaceae). *Molecular Biology and Evolution* 15 (5), 552-559

Citerne, H. L., M. Moller et Q. C. B. Cronk. 2000. Diversity of *cycloidea*-like genes in Gesneriaceae in relation to floral symmetry. *Annals of Botany* 86, 176-176

Clegg, M. T., M. P. Cummings et M. L. Durbin. 1997. The evolution of plant nuclear genes. *Proceedings of the National Academy of Sciences of the United States of America* 94 (15), 7791-7798

Coen, E. S., J. M. Romero, S. Doyle, R. Elliot, G. Murphy et R. Carpenter. 1990. *floricaula*: a homeotic gene required for flower development in *Antirrhinum majus*. *Cell* 63, 1311-1322

Dallas-Wang, Q., G. Jiang et F. M. Sladek. 1998. Avoiding false positives in colony PCR. *BioTechniques* 24 (4), 580-582

Darnell, J. E. et W. F. Doolittle. 1986. *Proceedings of the National Academy of Sciences of the United States of America* 83, 1271-1275

Denton, A. L., B. L. McConaughy et B. D. Hall. 1998. Usefulness of RNA polymerase II coding sequences for estimation of green plant phylogeny. *Molecular Biology and Evolution* 15 (8), 1082-1085

Devlin, P. F. et S. A. Kay. 2000. Flower arranging in *Arabidopsis*. *Science* 288 (5471), 1600-16002

Dibb, N. J. et A. J. Newman. 1989. Evidence that intron arose at proto-splice sites. *The Embo Journal* 8 (7), 2015-2021

Dickinson, W. C. 1981. The evolutionary relationships of the Leguminosae. *Dans* R. M. Polhill et P. H. Raven (eds), Advances in Legume Systematics, part 1. Royal Botanic Gardens, Kew, pp. 35-54.

Donoghue, M. J. et S. Mathews. 1998. Duplicate genes and the root of angiosperms, with an exemple using phytochrome sequence. *Molecular Phylogenetics and Evolution* 9 (3), 489-500

Doyle, J. J. 1992. Gene trees and species trees: molecular systematics as one character taxonomy. *Systematic Botany* 17 (1), 144-163

Doyle, J. J. 1995. DNA data and legume phylogeny: a progress report. *Dans* M. Crisp et J. J. Doyle (eds), Advances in Legume Systematics 7: Phylogeny. Royal Botanic Gardens, Kew, pp. 11-30.

Doyle, J. J., J. A. Chappill, C. D. Bailey et T. Kajita. 2000. Towards a comprehensive phylogeny of legumes: evidence from rbcL sequences and non-molecular data. *Dans* P. S. Herendeen et A. Bruneau (eds), Advances in Legume Systematics, part 9. Royal Botanic Gardens, Kew, pp. 1-20.

Doyle, J. J. et J. I. Davis. 1998. Homology in molecular phylogenetics: a parsimony perspective. *Dans* D. E. Soltis, P. S. Soltis et J. J. Doyle (eds), Molecular systematics of plants II. Kluwer Academic Publishers, Boston, pp. 101-132.

Doyle, J. J. et J. L. Doyle. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* 19 (1), 11-15

Doyle, J. J., J. L. Doyle, J. A. Ballenger, E. E. Dickson, T. Kajita et H. Ohashi. 1997. A phylogeny of the chloroplast gene *rbcL* in the Leguminosae: taxonomic correlations and insights into the evolution of nodulation. *American Journal of Botany* 84 (4), 541-554

Doyle, J. J., V. Kanazin et R. C. Shoemaker. 1996a. Phylogenetic utility of histone H3 intron sequences in the perennial relatives of soybean (*Glycine*: Leguminosae). *Molecular Phylogenetics and Evolution* 6 (3), 438-447

Doyle, J. J., V. Kanazin et R. C. Shoemaker. 1996b. Phylogenetic utility of histone H3 intron sequences in the perrenial relatives of soybean (*Glycine*: Leguminosae). *Molecular Phylogenetics and Evolution* 6 (3), 438-447

Drouin, G., F. Prat, M. Ell et G. D. P. Clarke. 1999. Detecting and characterizing gene conversion between multigene family members. *Molecular Biology and Evolution* 16 (10), 1369-1390

Durbin, M. L., B. McCaig et M. T. Clegg. 2000. Molecular evolution of the chalcone synthase multigene family in the morning glory genome. *Plant Molecular Biology* 42, 79-92

Emshwiller, E. et J. J. Doyle. 1999. Chloroplast-expressed glutamine synthetase (ncpGS): Potential utility for phylogenetic studies with an example from *Oxalis* (Oxalidaceae). *Molecular Phylogenetics and Evolution* 12 (3), 310-319

Evans, R. C., L. A. Alice, C. S. Campbell, E. A. Kellogg et T. Dickinson. 2000. The Granule-Bound Starch Synthase (GBSSI) gene in the Rosaceae: multiple loci and phylogenetic utility. *Molecular Phylogenetics and Evolution* 17 (3), 388-400

Farris, J. S., M. Källersö, A. G. Kluge et C. Bult. 1994. Testing significance of incongruence. *Cladistics* 10, 315-319

Federoff, N. 2000. Transposons and genome evolution in plants. *Proceedings of the National Academy of Sciences of the United States of America* 97 (13), 7002-7007

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* 27, 401-410

Filatov, D. A. et D. Charlesworth. 1999. DNA polymorphism, haplotypre structure and balancing selection in the *Leavenworthia PgiC* locus. *Genetics* 153, 1423-1434

Ford, V. S. et L. D. Gottlieb. 1999. Molecular characterization of *PgiC* in a tetraploid plant and its diploid relatives. *Evolution* 53 (4), 1060-1067

Forest, F. et A. Bruneau. 2000. Phylogenetic analysis, organisation, and molecular evolution of the nontranscribed spacer of 5S ribosomal RNA genes in *Corylus* (Betulaceae). *International Journal of Plant Science* 161 (5), 793-806

Friedlander, T. P., J. C. Regier, C. Mitter et D. L. Wagner. 1996. A nuclear gene for higher level phylogenetics: phosphoenolpyruvate carboxykinase tracks Mesozoic-age of divergences within Lepidoptera (Insecta). *Molecular Biology and Evolution* 13 (4), 594-604

Frohlich, M. W. et E. M. Meyerowitz. 1997. The search for flower homeotic gene homologs in basal angiosperms and gnetales: a potential new source for data on the evolutionary origin of flowers. *International Journal of Plant Science* 158 (Supplement 6), 131-142

Frohlich, M. W. et D. S. Parker. 2000. The mostly-male theory of flower evolutionary origin: from genes to fossils. *Systematic Botany* 25 (2), 155-170

Frugoli, J. A., M. A. McPeek, T. T. Thomas et C. R. McClung. 1998. Intron loss and gain during evolution of the catalase gene family in angiosperms. *Genetics* 149, 355-365

Galloway, G. L., R. L. Malmberg et R. A. Price. 1998. Phylogenetic utility of the nuclear gene arginine decarboxylase: an exemple from Brassicaceae. *Molecular Biology and Evolution* 15 (10), 1312-1320

Gaut, B. S., B. R. Morton, B. McCaig et M. T. Clegg. 1996. Substitution rate comparison between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. *Proceedings of the National Academy of Sciences of the United States of America* 93, 10274-10279

Gaut, B. S., A. S. Peek, B. R. Morton et M. T. Clegg. 1999. Patterns of genetic diversification within the *Adh* gene family in the grasses (Poaceae). *Molecular Biology and Evolution* 16 (8), 1086-1097

Gilbert, W. 1978. Why genes in pieces? *Nature* 271, 501

Gilbert, W., M. Marchionni et G. McKnight. 1986. On the antiquity of introns. *Cell* 46, 151-154

Giroux, M. J., M. Clancy, J. Baier, L. Ingham, M. Donald et L. C. Hannah. 1994. *De novo* synthesis of an intron by the maize transposable element *Dissociation*. *Proceedings of the National Academy of Sciences of the United States of America* 91, 12150-12154

Gottlieb, L. D. et V. S. Ford. 1997. A recently silenced, duplicate *PgiC* locus in *Clarkia*. *Molecular Biology and Evolution* 14 (2), 125-132

Gourlay, C. W., J. M. I. Hofer et T. H. N. Ellis. 2000. Pea compound leaf architecture is regulated by interactions among the genes *UNIFOLIATA*, *COCHLEATA*, *AFFILA*, and *TENDRIL-LESS*. *The Plant Cell* 12, 1279-1294

Graybeal, A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? *Systematic Biology* 47 (1), 9-17

Hasebe, M. et J. A. Banks. 1997. Evolution of MADS gene family in plants. *Dans* Katwatsuki et P. H. Raven (eds), Evolution and diversification of land plants. Springer-Verlag, Tokyo, pp. 179-197.

Helariutta, Y., M. Kotilainen, P. Elomaa, N. Kalkkinen, K. Bremer, T. H. Teeri et V. A. Albert. 1996. Duplication and functional divergence in the chalcone synthase gene family of Asteraceae: evolution with substrate change and catalytic simplification. *Proceedings of the National Academy of Sciences of the United States of America* 93, 9033-9038

Herendeen, P. S., W. L. Crepet et D. L. Dilcher. 1992. The fossil history of the Leguminosae: phylogenetic and biogeographic implications. *Dans* P. S.

Herendeen et D. L. Dilcher (eds), Advances in Legume Systematics 4: The Fossil Record. Royal Botanic Gardens, Kew, pp. 303-316.

Herendeen, P. S. et B. F. Jacobs. 2000. Fossil legumes from the middle Eocene (40.6 Ma) Mahenge flora of Singida, Tanzania. *American Journal of Botany* 87 (9), 1358-1366

Hofer, J., L. Turner, R. Hellens, M. Ambrose, P. Matthews, A. Michael et N. Ellis. 1997. Unifoliata regulates leaf and flower morphogenesis in pea. *Current Biology* 7, 581-587

Hofer, J. M. I. et T. H. N. Ellis. 1998. The genetic control of patterning in pea leaves. *Trends in Plant Science* 3 (11), 439-444

Hsieh, J. et A. Fire. 2000. Recognition and silencing of repeated DNA. *Annuals Reviews in Genetics* 34, 187-204

Huttley, G. A., M. L. Durbin, D. E. Glover et M. T. Clegg. 1997. Nucleotide polymorphism in the chalcone synthase-A locus and evolution of the chalcone synthase multigene family of common morning glory *Ipomea purpurea*. *Molecular Ecology* 6, 549-558

Innan, H., F. Tajima, R. Terauchi et N. T. Miyashita. 1996. Intragenic recombination in the *Adh* locus of the wild plant *Arabidopsis thaliana*. *Genetics* 143, 1761-1770

Iwamoto, M., M. Maekawa, A. Saito, H. Higo et K. Higo. 1998. Evolutionary relationship of plant catalase genes inferred from exon-intron structures: isozyme divergence after the separation of monocots and dicots. *Theoretical and Applied Genetics* 97, 9-19

Jensen, U. 1995. Serological legumin data and the phylogeny of the Ranunculaceae. *Plant Systematics and Evolution* Supplement 9, 217-227

Joly, S., L. Brouillet et A. Bruneau. 2001. Phylogenetic implications of the multiple losses of the mitochondrial coxII.i3 intron in the angiosperms. *International Journal of Plant Science* 162 (2), 359-373

Käss, E. et M. Wink. 1997. Phylogenetic relationships in the Papilionoideae (Family Leguminosae) based on nucleotide sequences of cpDNA (*rbc*L) and ncDNA (ITS 1 and 2). *Molecular Phylogenetics and Evolution* 8 (1), 65-88

Kawabe, A., K. Yamane et N. T. Miashita. 2000. DNA polymorphism, at the cytosolic phosphoglucose isomerase (*PgiC*) locus of wild plant *Arabidopsis thaliana*. *Genetics* 156, 1339-1347

Kelchner, S. A. 2000. The evolution of non-coding chloroplast DNA and its application in plant systematics. *Annals of the Missouri Botanical Garden* 87 (4), 482-498

Kellogg, E. A. 1996. Integrating genetics, phylogenetics and developmental biology. *Dans* B. Sobral (ed) The impact of plant molecular genetics. Birkhäuser, Boston, pp. 159-172.

Kellogg, E. A., R. Appels et R. J. Mason-Gamer. 1996. When genes tell different stories: the diploid genera of Triticaceae (Gramineae). *Systematic Biology* 21 (3), 321-347

Klitgaard, B. B. 1991. Ecuadorian *Brownea* and *Browneopsis* (Leguminosae-Caesalpinioideae): taxonomy, palynology, and morphology. *Nordic Journal of Botany* 11, 433-449

Koch, M. A., B. Haubold et T. Mitchell-Olds. 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis* and related genera (Brassicaceae). *Molecular Biology and Evolution* 17 (10), 1483-1498

Koch, M. A., B. Haubold et T. Mitchell-Olds. 2001. Molecular systematics of the Brassicaceae: evidence from coding plastid *matK* and nuclear *Chs* sequences. *American Journal of Botany* 88 (2), 534-544

Kramer, E. M., R. L. Dorit et V. F. Irish. 1998. Molecular evolution of genes controlling petal and stamen development: duplication and divergence within the *APETALA3* and *PISTILLATA* MADS-Box gene lineages. *Genetics* 149, 765-783

Kumar, S., K. Tamura, I. B. Jakobsen et M. Nei. 2001. MEGA: Molecular Evolutionary Genetics Analysis software. Ver. 2.0

Lang, J. et H. Fisher. 1995. Cloning, sequencing, and phylogenetic analysis of a legumin cDNA of *Hepatica nobilis* (*Ranunculaceae*). *Plant Systematics and Evolution* Supplement 9, 301-303

Lavin, M., E. Eshbaugh, J.-M. Hu, S. Mathews et R. A. Sharrock. 1998. Monophyletic subgroup of the tribe Milletiaea (Leguminoseae) as revealed by phytochrome nucleotide sequence data. *American Journal of Botany* 85 (3), 412-433

Lavin, M., R. T. Pennington, B. B. Klitgaard, J. I. Sprent, H. C. de Lima et P. E. Gasson. 2001. The dalbergioid legumes (Fabaceae): Delimitation af a pantropical monophyletic clade. *American Journal of Botany* 88 (3), 503-533

Lawton-Rauh, A., E. S. Buckler IV et M. D. Purugganan. 1999. Patterns of molecular evolution among paralogous floral homeotic loci. *Molecular Biology and Evolution* 16 (8), 1037-1045

Leitch, I. J., M. W. Chase et M. D. Bennett. 1998. Phylogenetic analysis of DNA C-values provides evidence for a small ancestral genome size in flowering plants. *Annals of Botany* 82 (Supplement A), 85-94

Lewin, B. 2000. *GENE VII*. Oxford University Press Inc., New York. 990 p.

Li, W. H. 1997. *Molecular evolution*. Sinauer Associates Inc., Sunderland. 487 p.

Lin, J.-Z., A. H. D. Brown et M. T. Clegg. 2001. Heterogeneous geographic patterns of nucleotide sequence diversity between two alcohol deshydrogenase genes in wild barley (*Hordeum vulgare* subspecies spontaneum). *Proceedings of the National Academy of Sciences of the United States of America* 98 (2), 531-536

Liu, F., D. Charlesworth et M. Kreitman. 1999. The effect of mating system differences on nucleotide diversity at the phosphoglucose isomerase locus in the plant genus *Leavenworthia*. *Genetics* 151, 343-357

Lodish, H., A. Birk, S. L. Zipursky, P. Matsudaira, D. Baltimore et J. Darnell. 1999. *Molecular cell biology*. W.H. Freeman and Company, New York. 1084 p.

Logsdon, J. M., M. G. Tyshenko, C. Dixon, J. D.-Jafari, V. K. Walker et J. D. Palmer. 1995. Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the intron-late theory. *Proceedings of the National Academy of Sciences of the United States of America* 92, 8507-8511

Luckow, M. P., J. White et A. Bruneau. 2000. Relationships among the basal genera of mimosoid legumes. *Dans* P. S. Herendeen et A. Bruneau (eds), Advances in Legume Systematics, part 9. Royal Botanic Gardens, Kew, pp. 165-180.

Luo, D., R. Carpenter, C. Vincent, L. Copsey et E. S. Coen. 1996. Origin of floral asymmetry in *Antirrhinum*. *Nature* 383 (6603), 794-799

Maddison, D. R. et W. P. Maddison. 2000. MacClade. Ver. 4.0. Sinauer Associates Inc., Sunderland.

Maleki, S. J., R. A. Kopper, D. S. Shin, C.-W. Park, C. M. Compandre, H. Sampson, A. W. Burks et G. A. Bannon. 2000. Structure of the major peanut allergen Ara h1 may protect IgE-binding epitope from degradation. *The Journal of Immunology* 164, 5844

Marchunk, D., M. Drumm, A. Saulino et F. S. Collins. 1990. Construction of T-vector, a rapid and general system for direct cloning of unmodified PCR products. *Nucleic Acids Research* 19 (5), 1154

Martin, W., D. Lydiate, H. Brinkmann, G. Forkmann, H. Saedler et R. Cerff. 1993. Molecular phylogenies in angiosperm evolution. *Molecular Biology and Evolution* 10 (1), 140-162

Martin, W., B. Stoebe, V. Goremykin, S. Hansmann, M. Hasegawa et K. V. Kowallik. 1998. Gene transfer to the nucleus and the evolution of chloroplast. *Nature* 393, 162-165

Mason-Gamer, R. J., C. F. Weil et E. A. Kellogg. 1998. Granule-Bound Starch Synthase: structure, function, and phylogenetic utility. *Molecular Biology and Evolution* 15 (12), 1658-1673

Masterson, J. 1994. Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science* 264 (5157), 421-424

Mathews, S. et M. J. Donoghue. 1999. The root of angiosperms phylogeny inferred from duplicate phytochrome genes. *Science* 286 (5441), 947-950

Mathews, S., M. Lavin et R. A. Sharrock. 1995. Evolution of the phytochrome gene family and its utility for phylogenetic analysis in angiosperms. *Annals of the Missouri Botanical Garden* 82, 296-361

Mathews, S., R. C. Tsai et E. A. Kellogg. 2000. Phylogenetic structure in the grass family (Poaceae): evidence from the nuclear gene phytochrome B. *American Journal of Botany* 87 (1), 96-106

McDade, L. 1995. Hybridization and phylogenetics. *Dans* P. C. Hoch et A. G. Stephenson (eds), Experimental and molecular approaches to plant biosystematics. Monographs in Systematic Botany from the Missouri Botanical Garden, pp. 305-331.

Miller, R. E., M. D. Rausher et P. D. Manos. 1999. Phylogenetic systematics of *Ipomea* (Convolvulaceae) based on ITS and *WAXY* sequences. *Systematic Botany* 24 (2), 209-227

Miyashita, N. T. 2001. DNA variation in the 5' upstream region of the *Adh* locus of the wild plants *Arabidopsis thaliana* and *Arabis gemmifera*. *Molecular Biology and Evolution* 18 (2), 164-171

Moniz de Sa, M. et G. Drouin. 1996. Phylogeny and substitution rates of angiosperms actin genes. *Molecular Biology and Evolution* 13 (9), 1198-1212

National Center for Biotechnology Information. 2001a. *Entrez Genome*. [En ligne]. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome

National Center for Biotechnology Information. 2001b. *GenBank*. [En ligne]. http://www.ncbi.nlm.nih.gov/BLAST/

Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlag, New York. 160 p.

Organelle Genome Megasequencing Program. 2001. *GOBASE*. [En ligne]. http://megasun.bch.umontreal.ca/gobase (mars 2001).

Oxelman, B. et B. Bremer. 2000. Discovery of paralogous nuclear gene sequences coding for the second largest subunit of RNA polymerase II (*RPB2*) and their phylogenetic utility in Gentianales of the Asterid. *Molecular Biology and Evolution* 17 (8), 1131-1145

Page, R. D. M. et M. A. Charleston. 1997. From gene to organismal phylogeny : reconciled trees and the gene tree/species tree problem. *Molecular Phylogenetics and Evolution* 7, 231-240

Palmer, J. D., K. L. Adams, Y. Cho, C. L. Parkinson, Y.-L. Qiu et K. Song. 2000. Dynamic and evolution of plant mitochondrial genomes: mobile genes

and introns and highly variable mutation rates. *Proceedings of the National Academy of Sciences of the United States of America* 97 (13), 6960-6966

Palmer, J. D. et J. M. Logdson. 1991. The recent origins of introns. *Current Opinion in Genetics and Development* 1, 470-477

Parcy, F., O. Nilsson, M. A. Bush, I. Lee et D. Weigel. 1998. A genetic framework for floral patterning. *Nature* 385 (6702), 561-566

Patthy, L. 1985. Evolution of the proteases of blood coagulation and fibrinolysis by assembly from modules. *Cell* 41, 657-663

Pelaz, S., G. S. Ditta, E. Baumann, E. Wisman et M. F. Yanofsky. 2000. B and C floral organ identity function require *SEPALLATA* MADS-box genes. *Nature* 405 (6783), 200-203

Pennington, R. T., B. B. Klitgaard, H. Ireland et M. Lavin. 2000. New insights into floral evolution of basal Papilionoideae from molecular phylogenies. *Dans* P. S. Herendeen et A. Bruneau (eds), Advances in Legume Systematics, part 9. Royal Botanic Gardens, Kew, pp. 233-248.

Perry, D. O. et G. R. Furnier. 1996. *Pinus banksiana* has at least seven expressed alcohol dehydrogenase genes in two linked groups. *Proceedings of the National Academy of Sciences of the United States of America* 93, 13020-13023

Pohlmeyer, K., B. K. Paap, J. Soll et N. Wedel. 1996. CP12: a small nuclear-encoded chloroplast protein provides novel insight into higher-plant GAPDH evolution. *Plant Molecular Biology* 32 (5), 969-978

Polhill, R. M. 1994. Complete synopsis of legume genera. *Dans* F. A. Bisby, J. Buckingham et J. B. Harborne (eds), Phytochemical dictionary of the Leguminosae. Chapman and Hall, London, pp. xlix-liv.

Polhill, R. M., P. H. Raven et C. H. Stirton. 1981. Evolution and systematics of the Leguminosae. *Dans* R. M. Polhill et P. H. Raven (eds), Advances in Legume Systematics, part 1. Royal Botanic Gardens, Kew, pp. 1-26.

Purugganan, M. D. 1998. The molecular evolution of development. *Bioessays* 20 (9), 700-711

Raven, P. H. et R. M. Polhill. 1981. Biogeography of the Leguminosae. *Dans* R. M. Polhill et P. H. Raven (eds), Advances in Legume Systematics, part 1. Royal Botanic Gardens, Kew, pp. 27-34.

Riechmann, J. L. et E. M. Meyerowitz. 1997. MADS domain proteins in plant development. *Biological Chemistry* 378 (10), 1079-1101

Rieseberg, L. H. et J. D. Morefield. 1995. Character expression, phylogenetic reconstruction, and the detection of reticulate evolution. *Dans* P. C. Hoch et A. G. Stephenson (eds), Experimental and molecular approaches to plant biosystematics. Monographs in Systematic Botany from the Missouri Botanical Garden, pp. 333-353.

Rogers, J. 1985. Exon shuffling and intron insertion in serine proteases genes. *Nature* 315, 458-459

Rokas, A. et P. W. H. Holland. 2000. Rare genomic changes as a tool for phylogenetics. *Trends in Ecology and Evolution* 15 (11), 454-459

Saenz de Miera, L. E. et M. Pérez de la Vega. 1998. A comparativ study of vicilin genes in *Lens*: Negative evidence of concerted evolution. *Molecular Biology and Evolution* 15 (3), 303-311

Sambrook, J., E. F. Fritsch et T. Maniatis. 1989. *Molecular cloning: a laboratory manual*. Cold Spring Harbor Laboratory, New York. p.

Sanderson, M. J. et J. J. Doyle. 1994. Reconstruction of organismal and gene phylogenies from data on multigene families: concerted evolution, homoplasy, and confidence. *Systematic Biology* 41 (1), 4-17

Sang, T. et D. Zhang. 1999. Reconstructing hybrid speciation using sequences of low copy nuclear genes: hybrid origins of five *Paeonia* species based on *Adh* gene phylogeny. *Systematic botany* 24 (2), 148-163

Sato, S., Y. Nakamura, T. Kaneko, E. Asamizu et S. Tabata. 1999. Complete structure of the chloroplast genome of *Arabidopsis thaliana*. *DNA Research* 6, 283-290

Savolainen, O., C. H. Langley, B. P. Lazzaro et H. Fréville. 2000. Contrasting patterns of nucleotide polymorphism at the alcohol dehydrogenase locus in the outcrossing *Arabidopsis lyrata* and the selfing *Arabidopsis thaliana*. *Molecular Biology and Evolution* 17 (4), 645-655

Schmidt, T. et J.-S. Heslop-Harrison. 1998. Genomes, genes and junk: the large scale organization of plant chromosomes. *Trends in Plant Science* 3 (5), 195-199

Schultz, E. A. et G. W. Haughn. 1991. *LEAFY*, a homeotic gene that regulates inflorescence development in *Arabidopsis*. *The Plant Cell* 3, 771-781

Shannon, S. et D. R. Meeks-Wagner. 1993. Genetic interaction that regulates inflorescence development in *Arabidopsis*. *The Plant Cell* 5 (6), 639-655

Shu, G., W. Amaral, L. C. Hileman et D. A. Baum. 2000. *LEAFY* and the evolution of rosette flowering in violet cress (*Jonopsidium acaule*, Brassicaceae). *American Journal of Botany* 87 (5), 634-641

Siddal, M. E. 1998. Success of parsimony in the four-taxon case - Long branch repulsion by likelihood in the Farris zone. *Cladistics* 14 (3), 209

Sidow, A. et W. K. Thomas. 1994. A molecular evolutionary framework for eukaryotic model organisms. *Current Biology* 4 (7), 596-603

Simmons, M. P., C. C. Clevinger, V. Savolainen, R. H. Archer, S. Mathews et J. J. Doyle. 2001. Phylogeny of Celastraceae inferred from phytochrome B gene sequence and morphology. *American Journal of Botany* 88 (2), 313-325

Simpson, G. G. et W. Filipowicz. 1996. Splicing of precursor to mRNA in higher plants: mechanism, regulation, and subnuclear organisation of the spliceosomal machinery. *Plant Molecular Biology* 32, 1-41

Singer, S., J. Sollinger, S. Maki, J. Fishback, B. Short, C. Reinke, J. Fick, L. Cox, A. McCall et H. Mullen. 1999. Inflorescence architecture: a developmental approach. *The Botanical Review* 65 (4), 385-410

Small, R. L., J. A. Ryburn, R. C. Cronn, T. Seelanan et J. F. Wendel. 1998. The tortoise and the hare: choosing between noncoding plastome and nuclear *Adh* sequences for phylogeny reconstruction in a recently diverged plant group. *American Journal of Botany* 85 (9), 1301-1315

Small, R. L. et J. F. Wendel. 2000. Copy number lability and evolutionary dynamics of the *Adh* gene family in diploid and tetraplod cotton (*Gossypium*). *Genetics* 155, 1913-1926

Souer, E., A. van der Krol, D. Kloos, C. Spelt, M. Bliek et J. Mol. 1998. Genetic control of branching pattern and floral identity during *Petunia* inflorescence development. *Development* 125, 733-742

Swofford, D. L. 2000. PAUP* phylogenetic analysis using parsimony (and other methods),. Ver. 4b8. Sinauer, Sunderland, Massachusset.

Swofford, D. L., G. J. Olsen, P. J. Waddell et D. M. Hillis. 1996. Phylogenetic Inference. *Dans* D. M. Hillis, C. Moritz et B. K. Mable (eds), Molecular Systematics. Sinauer Associates, Inc, Sunderland, pp. 407-514.

Theissen, G. 2000. Plant Breedings: FLO-like meristem identity genes: from basic science to crop plant design. Progress in Botany. Springer-Verlag, pp. 167-183.

Thompson, J. et F. Jeanmougin. 2000. Clustal X multiple sequence alignement program. Ver. 1.81

Tucker, S. 1987. Floral initiation and development in legumes. *Dans* C. H. Stirton (ed) Advances in Legume Systematics, part 3. Royal Botanic Gardens, Kew, pp. 183-239.

Tucker, S. 2001. The ontogenetic basis for missing petals in *Crudia* (Leguminosae: Caesalpinioideae: Detarieae). *International Journal of Plant Science* 162 (1), 83-89

Tucker, S. C. 2000a. Evolutionary loss of sepals and/or petals in detarioid legume taxa *Aphanocalyx Brachystegia*, and *Monopetalanthus* (Leguminosae: Caesalpinioideae). *American Journal of Botany* 87 (5), 608-624

Tucker, S. C. 2000b. Floral development and homeosis in *Saraca* (Leguminoseae: Caesalpinioideae: Detarieae). *International Journal of Plant Science* 161 (4), 537-549

Tucker, S. C. et A. W. Douglas. 1994. Ontogenetic evidence and phylogenetic relationships among basal taxa of legumes. *Dans* I. K. Ferguson et S. Tucker (eds), Advances in Legume Systematics 6: Structural Botany. Royal Botanic Gardens, Kew, pp. 11-32.

Tyshenko, M. G. et V. K. Walker. 1997. Towards a reconciliation of the intron early or late views: triosephosphate isomerase genes from insects. *Biochimica et Biophysica Acta* 1353, 131-136

Velazquez, D. 1992. Recognition of *Brownea coccinea* Jacq. subspecies *capitella* (Leguminosae). *Novon* 2, 173-175

Venkatesh, B., Y. Ning et S. Brenner. 1999. Late changes in spliceosomal introns define clades in vertebrate evolution. *Proceedings of the National Academy of Sciences of the United States of America* 96, 10267-10271

Vieira, C. P., J. Vieira et D. Charlesworth. 1999. Evolution of the cycloidea gene family in *Antirrhinum* and *Misopates*. *Molecular Biology and Evolution* 16 (11), 1474-1483

Vision, T. J., D. G. Brown et S. D. Tanksley. 2000. The origins of genomic duplications in *Arabidopsis*. *Science* 290 (5499), 2114-2117

Wagner, A. 1998. The fate of duplicated genes: loss or new function? *BioEssays* 20 (785-788)

Walker, E. L., N. F. Weeden, C. B. Taylor, P. Green et G. M. Coruzzi. 1995. Molecular evolution of duplicate copies of genes encoding cytosolic glutamine synthase in *Pisum sativum*. *Plant Molecular Biology* 29 (6), 1111-1125

Wang, X.-Q., D. C. Tank et T. Sang. 2000. Phylogeny and divergence times in Pinaceae: evidence from three genomes. *Molecular Biology and Evolution* 17 (5), 773-781

Watson, M. W., W. Yu, G. L. Galloway et R. L. Malmberg. 1997. Isolation and characterization of a second arginine decarboxylase cDNA from *Arabidopsis*. *Plant Physiology* 114, 1569

Wendel, J. F. et J. J. Doyle. 1998. Phylogenetic incongruence: window into genome history and molecular evolution. *Dans* D. E. Soltis, P. S. Soltis et J. J. Doyle (eds), Molecular systematics of plants II. Kluwer Academic Publishers, Boston, pp. 265-296.

Whitlock, B. A. et D. A. Baum. 1999. Phylogenetic relationships of *Theobroma* and *Herrania* (Sterculiaceae) based on sequences of the nuclear gene *vicilin*. *Systematic Botany* 24 (2), 128-138

Wolfe, K. H., P. M. Sharp et W. H. Li. 1989. Rates of synonymous substitution in plant genes. *Journal of Molecular evolution* 29, 208-211

Xu, S. 2000. Phylogenetic analysis under reticulate evolution. *Molecular Biology and Evolution* 17 (6), 897-907

Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology and Evolution* 11, 367-372

Yang, Z. et S. Kumar. 1996. Approximate methods of estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. *Molecular Biology and Evolution* 13 (5), 650-659