Plant Syst. Evol. (2004)
DOI 10.1007/s00606-004-0223-2

Plant Systematics and Evolution
Printed in Austria

Phylogenetic analysis of *Theobroma* (Sterculiaceae) based on Kunitz-like trypsin inhibitor sequences

C. R. Sousa Silva and **A. Figueira**

Laboratório de Melhoramento de Plantas, Centro de Energia Nuclear na Agricultura, Universidade de São Paulo, Piracicaba, São Paulo, Brazil

Received March 18, 2004; accepted June 15, 2004
Published online: December 22, 2004

Abstract. Trypsin inhibitor gene sequences were used to investigate the phylogenetic relationships among *Theobroma* and *Herrania* species, considered as sister-groups, with particular interest on the monophyly and infra-generic relationships of *Theobroma*. The presumed amino acid sequences of 23 analyzed samples, from 11 *Theobroma* and three *Herrania* species, comprising all sections from both genera, demonstrated a high similarity with a previously characterized *T. cacao* Kunitz trypsin inhibitor. The trypsin inhibitor gene accumulated mutations at faster rate than prior analyzed nuclear or chloroplastic genes. None of the sequences presented introns. The phylogeny of the trypsin inhibitor sequences was congruent with the phylogenetic hypotheses of the *Theobroma* and *Herrania* species based on morphology. The monophyly of *Theobroma* was not strongly supported, corroborating previously described absence of obvious synapomorphies for *Theobroma*. The species grouped consistently according to genus and section. The monophyly of all *Theobroma* sections was supported, except for section *Glossopetalum*, which was paraphyletic to section *Andropetalum*. Evidences sustain that *T. mammosum* may be included into section *Glossopetalum*. The potential use of trypsin inhibitor gene sequences in phylogenetic studies of *Theobroma* was demonstrated.

Key words: Cacao, cocoa, *Herrania*, Kunitz, Malvaceae s.l., *Theobroma*; trypsin inhibitor.

The genus *Theobroma* L. consists of small understory trees, which occurs natively in Neotropical lowland rainforests, extending from the Amazon basin through Southern Mexico (Cuatrecasas 1964). *Theobroma* is the genus of most relevance of the former Sterculiaceae (Purseglove 1968), because of the economic importance of *T. cacao* L. (cacao), the unique source of cocoa butter and solids for the chocolate and cosmetic industries. Recent phylogenetic studies based on combined analysis of plastid *atp*B, *ndh*F and *rbc*L sequences; morphology; and chemical characteristics were used to propose the inclusion of Sterculiaceae into a broadly defined Malvaceae (Alverson et al. 1999, Bayer et al. 1999, Judd and Manchester 1997, Whitlock et al. 2001).

Theobroma contains 22 species (Table 1) classified into six sections (Cuatrecasas 1964), which have a great importance as gene reservoir for cacao improvement, besides having an immense potential as new crops. The phylogenetic relationships of *Theobroma* had been defined by the classical method of comparative

Table 1. Species and sections from *Theobroma* (Cuatrecasas 1964) and *Herrania* (Schultes 1958), indicating the analyzed species by Asterisks

Genus	Sections	Species
Theobroma	*Andropetalum*	**T. mammosum* Cuatr. & Leon
	Glossopetalum	**T. angustifolium* Moçiño & Sessé
		T. canumanense Pires & Fróes
		T. chocoense Cuatr.
		T. cirmolinae Cuatr.
		**T. grandiflorum* (Willd. ex Spreng.) Schum.
		T. hylaeum Cuatr.
		T. nemorale Cuatr.
		**T. obovatum* Klotzsch ex Bernoulli
		**T. simiarum* Donn. Smith.
		T. sinuosum Pavón ex Hubber
		T. stipulatum Cuatr.
		**T. subincanum* Mart.
	Oreanthes	*T. bernouillii* Pittier
		T. glaucum Karst
		**T. speciosum* Willd.
		**T. sylvestre* Mart
		T. velutinum Benoist
	Rhytidocarpus	**T. bicolor* Humb. & Bonpl.
	Telmatocarpus	*T. gileri* Cuatr.
		**T. microcarpum* Mart.
	Theobroma	**T. cacao* L.
Herrania	*Subcymbicalyx*	*H. balaënsis* Preuss
		H. breviligulata Schultes
		H. camargoana Schultes
		H. cuatrecasana García-Barriga
		H. dugandii García-Barriga
		H. kanukuensis Schultes
		H. kofanorum Schultes
		H. laciniifolia Goudot
		H. lemniscata (Schomb.) Schultes
		**H. mariae* (Mart.) Decaisne ex Goudot
		H. nitida (Poepp.) Schultes
		H. nycterodendron Schultes
		H. pulcherrima Goudot
		H. tomentella Schultes
	Herrania	**H. albiflora* Goudot
		H. purpurea (Pitt.) Schultes
		H. umbratica Schultes

morphology, based mainly on pod and floral structure, and relevant vegetative characters (mode of germination; growth type; presence of hairs in young leaves; and branching system) (Cuatrecasas 1964). Section *Glossopetalum* was considered the most ancestral of the genus, while sections *Rhytidocarpus*, *Oreanthes* and *Theobroma* were considered more derived, with section *Theobroma* exhibiting the most apomorphic characters (Cuat-

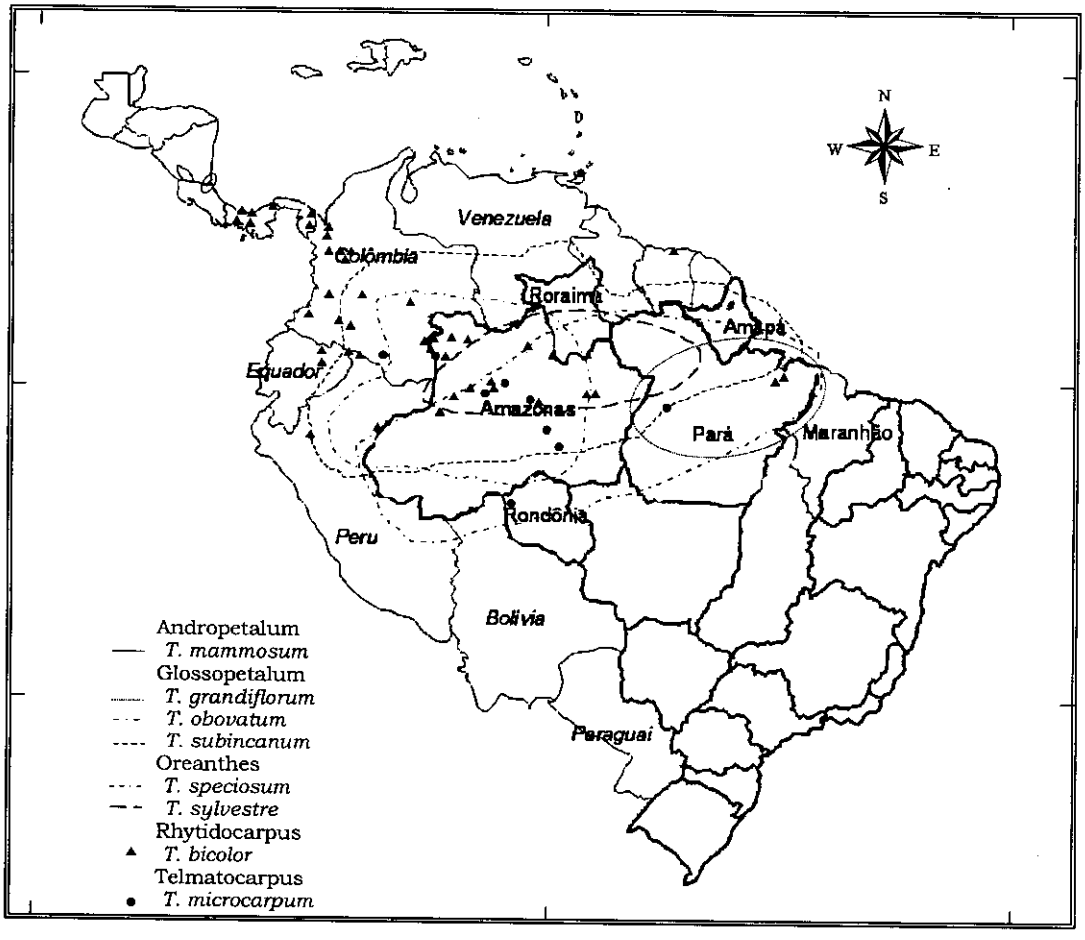

Fig. 1. Area of natural distribution of Brazilian species of *Theobroma* (sections *Glossopetalum* and *Oreanthes*); site of collection of *T. bicolor* (section *Rhytidocarpus*) and *T. microcarpum* (section *Telmatocarpus*) based on information of herbaria; and area of distribution of section *Andropetalum* (Central America) [adapted from Cuatrecasas 1964]

recasas 1964). Section *Telmatocarpus* displays some plesiomorphic characters, similar to *Glossopetalum*, while exhibiting other more apomorphic characters, similar to section *Theobroma* (e.g. glabrous leaves; partially woody pericarp) (Cuatrecasas 1964).

With the exception of the cultivated species (*T. cacao*, *T. bicolor*, and *T. grandiflorum*), most species of *Theobroma* have restricted natural geographical distributions, with major separation between species occurring at the region east of the Andes (Amazon) and at the Pacific side (Fig. 1). Three species (*T. simiarum*, *T. angustifolium*, and *T. mammosum*) are restricted to Central America (Cuatrecasas 1964). The uplift of the Andes might have separated previously spread *Theobroma* populations. Representative species from all sections (*T. grandiflorum*, *T. obovatum*, *T. subincanun*, *T. speciosum*, *T. sylvestre*, *T. microcarpum*, *T. bicolor*, *T. cacao*, *T. glaucum*, and *T. canumanense*), except for section *Andropetalum*, occur natively in the Brazilian Amazon (Table 1; Fig. 1) (Cuatrecasas 1964).

The genus *Herrania* Goudot is morphologically similar and closely related to *Theobroma* (Cuatrecasas 1964), being distinguished by a few morphological traits, such as the

unbranched trunk; the compound palmate leaves; and the long petal-lamina, exceeding the length of the petal-hood (Schultes 1958). Before the revision of Schultes (1958), *Herrania* was considered a section of *Theobroma*. *Herrania* has a similar geographic distribution to *Theobroma*, with only one species occurring in Central America [*H. purpurea* (Pitt.) Schultes], while the other 16 are limited to South America on both sides of the Andes (Schultes 1958). The *Herrania* species are classified into two sections: (1) section *Subcymbicalyx*, containing the 14 species with more plesiomorphic characteristics and a subcymbiform calyx; and (2) section *Herrania*, grouping three species with apomorphic characteristics and a pateliform calyx (Table 1) (Schultes 1958).

The close relationship between *Herrania* and *Theobroma* was supported by the cross compatibility and seed viability between *H. mariae* and other *Theobroma* species, including *T. cacao* (Addison and Tavares 1951), and by the lack of separation between both genera based on phenetic analysis of rDNA polymorphism (Figueira et al. 1994). Based on parsimony analysis of the vicilin gene, *Herrania* and *Theobroma* were considered monophyletic and sister genera, but only the monophyly of *Herrania* was strongly supported (Whitlock and Baum 1999).

Sequences of the trypsin inhibitor gene were here used to investigate the phylogenetic relationships among *Theobroma* and *Herrania* species. Seeds of *T. cacao* contain 15 to 20% of protein in dry weight, composed mainly (52% of total protein) by an albumin fraction (Spencer and Hodge 1991, Voigt and Biehl 1993). The 21 kDa trypsin inhibitor is the major constituent of the cacao seed albumin fraction, and its cDNA had been characterized, with the conceptually translated peptide presenting a high similarity with the soybean Kunitz trypsin inhibitor (Spencer and Hodge 1991, Tai et al. 1991). Proteinase inhibitory activity was later demonstrated for a protein extract from cacao seeds (Dodo et al. 1992). Protein extracts from seeds of the Brazilian *Theobroma* species exhibited major peptide bands of similar size (ca. 21 kDa) (Silva et al. 2001) and displayed trypsin inhibitory activity (unpublished results).

The objectives of this work were: 1. to evaluate the phylogeny of the 21 kDa Kunitz-like trypsin inhibitor gene sequences from the genera *Theobroma* and *Herrania*, evaluating its usefulness for phylogenetic studies; 2. to test the monophyly of the genus *Theobroma*; 3. to evaluate the phylogenetic relationships among the *Theobroma* sections; 4. to establish the phylogenetic relationships among the *Theobroma* species from the Brazilian Amazon and Central America.

Materials and methods

Taxonomic sampling. Eight Brazilian *Theobroma* species (*T. grandiflorum*, *T. subincanum*, *T. bicolor*, *T. sylvestre*, *T. obovatum*, *T. microcarpum*, *T. speciosum*, and *T. cacao*) plus *H. mariae* were sampled from the *Theobroma* collections "Addison O'Neill" at "Embrapa Amazônia Oriental", Belém, Pará, (1°20'S; 48°30'W); and "Basil Bartley" of the "Comissão Executiva do Plano da Lavoura Cacaueira - Ceplac", located at Marituba, PA (1°12'S; 49°30'W). Leaves were collected in the field, washed in tap and distilled water; blotted dry, and were stored frozen until DNA extraction. DNA samples from *T. angustifolium*, *T. mammosum*, *T. simiarum*, *Herrania albiflora*, and *Herrania* sp., derived from trees maintained at the germplasm collection of CATIE, Turrialba, Costa Rica, were obtained as described by Figueira et al. (1994).

DNA extraction. Frozen leaves were ground in liquid nitrogen, and DNA was extracted using a protocol adapted from Doyle and Doyle (1990). DNA concentration was determined spectrophotometrically.

Amplification of sequence homologues to the *T. cacao* seed trypsin inhibitor gene. Primers for the 21 kDa trypsin inhibitor gene were designed based on the 666 bp cDNA coding sequence from *T. cacao* (GenBank X56509.1) published by McHenry and Fritz (1992), to amplify a fragment ca. 465 bp. The amplification reactions contained 0.1 μM of each primer (tryp-for = 5' CTG TGC

TTG ACA CTG ATG GTG 3'; tryp-rev = 5' NNN TTC CAA TAT CGC TGC 3') and 20 ng of genomic DNA, and were conducted on a GeneAmp 9600 thermocycler (Applied Biosystems, Foster City, CA, EUA) programmed for 35 cycles of 94°C for 30 s; 40 s at 45°C, and 60 s at 72 °C, ending with a 7 min extension at 72°C. The products were analyzed in 2% agarose in TAE, ran at 6 V cm^{-1}.

Cloning and sequencing trypsin inhibitor gene homologues. The amplified fragments from all the *Theobroma* and *Herrania* species were cloned in pGEM-T (Promega, Madison, WI, USA). The plasmids were purified using Wizard-Plus Miniprep DNA Purification System (Promega). Positive clones were confirmed by digestion with *BstzI* or by amplification. Purified plasmids were sequenced in both directions in a 377 ABI automatic DNA analyzer (Applied Biosystems). Cloned sequences were deposited in GenBank (http://www.ncbi.nlm.nih.gov/) under accession numbers AF356602-AF356626.

Phylogenetic analyses. Nucleotide sequences of the cloned fragments were compared with the available *T. cacao* trypsin inhibitor cDNA (GenBank X56509.1) using BLASTn (Altschul et al. 1997). The sequences were conceptually translated into amino acids using EXPASY Translate tool (http://www.expasy.ch/tools/dna.html). The identification of amino acid functional substitution was conducted using PSI-BLAST (Altschul et al. 1997). Nucleotide and amino acid sequences were aligned using ClustalX 1.8 (Thompson et al. 1997) and visually. Parsimony analyses of the nucleotide and amino acid sequences were conducted using the software PAUP 4.0b10 (Swofford 1998). All searches for most parsimonious trees used heuristic searches with TBR and MULPARS and 100 random addition sequence replicates. The bootstrap values were estimated based on 1,000 replications with the same strategy used for the search for parsimony. The Decay Index (DI) (Bremer 1988) for each cluster was determined using the program SEPAL 1.4 (Salisbury 2000). Additionally, two distinct weighing schemes were tested in the parsimony analysis: (a) attributing a character weight 2:1:1 for the first, second and third codon position; (b) transition:transversion (1:2 stepmatrix). The *p*-distances between the sequences were estimated using PAUP 4.0b10 (Swofford 1998).

Results

Amplification, cloning and sequencing the seed 21 kDa trypsin inhibitor putative homologues. The proposed primers for the trypsin inhibitor gene successfully amplified a single fragment of the expected size (~465 bp) from all the *Theobroma* and *Herrania* species analyzed, with no apparent polymorphism under the standard gel conditions employed. The amplified fragments from the various species were cloned into pGEM-T vector and sequenced. It was attempted to sequence more than one clone per individual and more than one individual per species. One clone was obtained from *T. microcarpum*; 1 from *T. bicolor*; 1 from *T. subincanum*; 2 from *T. obovatum*; 3 from *T. sylvestre*; 1 from *T. speciosum*; 1 from *T. cacao*; 3 from *T. grandiflorum*; 1 from *T. angustifolium*; 3 from *T. mammosum*; 1 from *T. simiarum*; 1 from *H. albiflora*; 1 from *Herrania* sp.; and 3 from *H. mariae*. Sequencing was conducted on both directions for all clones.

Sequences characteristics. The multiple alignment of the 666 bp cDNA of the *T. cacao* trypsin inhibitor (GenBank X56509.1) with the nucleotide sequences from *Theobroma* and *Herrania* species had 465 characters (nucleotide and/or deletions), from which 295 were constant, 93 were variable and parsimonious informative, while 77 were variable but not informative. None of the nucleotide sequences presented introns. The *Herrania* nucleotide sequences showed 92% of identity with the original *T. cacao* cDNA sequence, while the *Theobroma* species showed on average 91.4% of identity. The *T. bicolor* clone 5-3 presented the lowest nucleotide identity with the *T. cacao* cDNA sequence (82%), while clone 34 from *T. cacao* presented 100% identity, as expected. The nucleotide sequences derived from the species of section *Oreanthes* (*T. speciosum* and *T. sylvestre*) presented a deletion of three bases in comparison to *T. cacao*. This deletion caused the loss of a Leucine, corresponding to the position 80 in the cacao 21 kDa trypsin

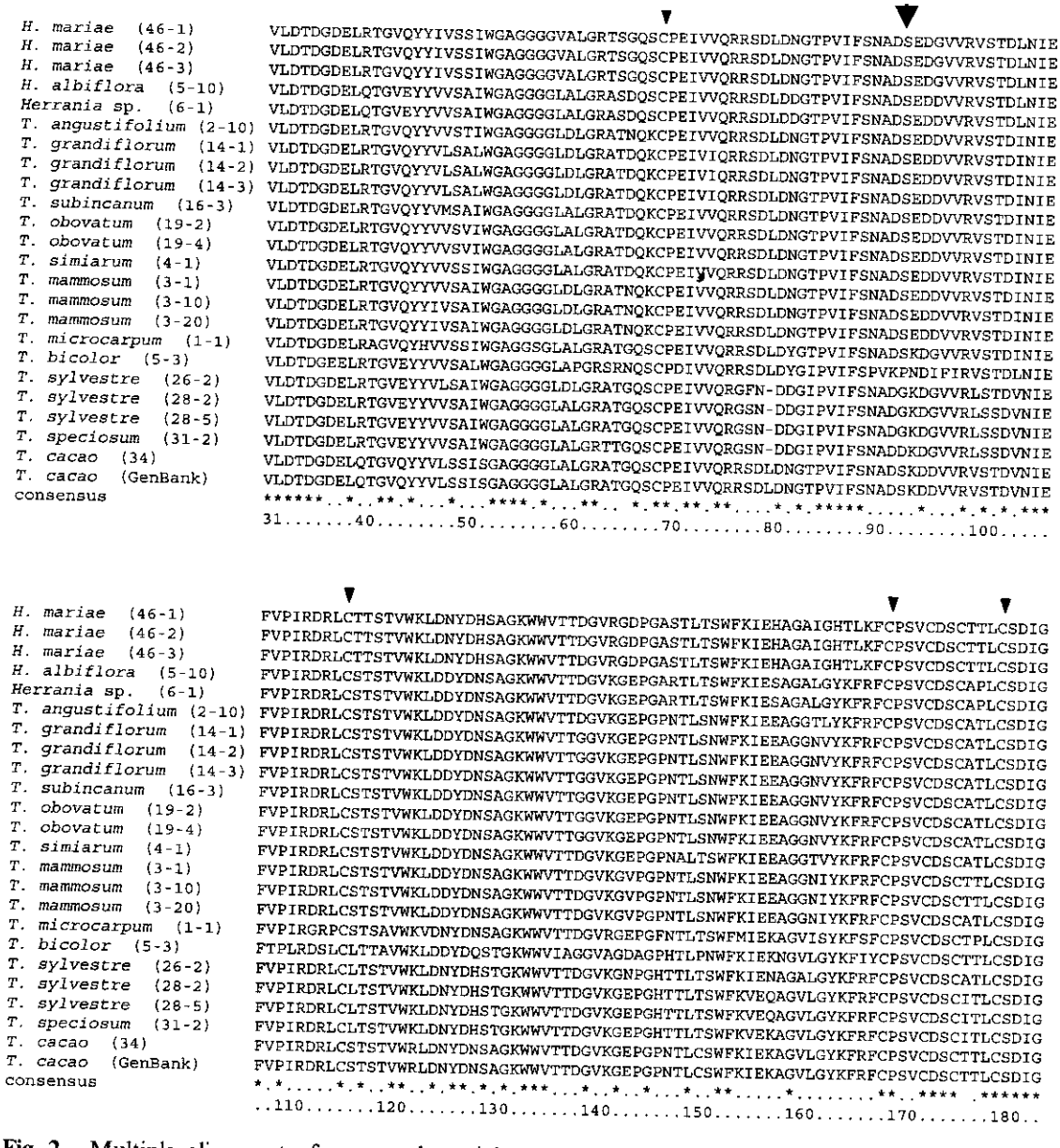

Fig. 2. Multiple alignment of presumed partial amino acid sequences of the trypsin inhibitor gene from *Theobroma* and *Herrania* species, indicating putative reactive site. Identical residues are identified by star, and similar residues by a dot at the consensus line. Residue numbering (ruler line) refers to the complete Kunitz-like trypsin inhibitor protein of *Theobroma cacao* (GenBank X56509.1). The position of the reactive site is indicated by an arrow, and cysteine residues responsible for the interchain disulfide bridges are marked (▼). The numbers between parentheses refer to plasmid clone number

inhibitor protein (Spencer and Hodge 1991, Tai et al. 1991; Fig. 2).

The uncorrected pairwise distance (*p*-distance) was calculated using a single nucleotide sequence representative from *Theobroma* (*T. cacao*) and from *Herrania* (*Herrania* sp.). The *p*-distance between both genera was 7.9%

The multiple alignment of the presumed amino acid sequences of the putative trypsin inhibitor from the *Theobroma* and *Herrania*

species (Fig. 2) resulted in 154 amino acid residues, with 73 constant characters, 48 variable and parsimonious informative characters, while 33 were variable but not informative. The average similarity between the presumed amino acid sequence of the *Herrania* species and the *T. cacao* trypsin inhibitor was 93%. *Herrania mariae* presented the lowest amino acid identity (83%) and the largest functional amino acid substitution (10%) from the genus *Herrania*. For the *Theobroma* species, the average amino acid sequence similarity with *T. cacao* was 91.7%, and the average sequence identity was 86.5%. *Theobroma simiarum* presented the highest identity with the cacao amino acid sequence (89%), with the lowest functional substitution of residues (4%), while clone 5-3 of *T. bicolor* had the lowest amino acid sequence identity (70%) with more functional substitutions (11%).

Parsimony analyses of nucleotide sequences. The analyses of the 23 resulting nucleotide sequences using simple parsimony, with equal character weight and gaps treated as missing data, resulted in sixteen equally parsimonious trees. The strict consensus of these trees resulted in a tree of 273 steps [Consistency index (CI) = 0.7839; Homoplasy index (HI) = 0.2161; CI excluding uninformative characters = 0.6828; HI excluding uninformative characters = 0.3172; Retention index (RI) = 0.8628; Rescaled consistency index (RC) = 0.6763]. The strict consensus is shown in Fig. 3 with bootstrap values and decay index values. One of the most parsimonious trees is shown with branches length in Fig. 4.

The monophyly of *Theobroma* was not strongly supported (bootstrap = 55%; D.I. = 2) (Fig. 3). The monophyly of *Theobroma* was supported by 7 characters: character 105 (C.I. = 1.00, change from T to A); 209 (C.I. = 1.00, change from T to C); 222 (C.I. = 0.50, change C to A); 351 (C.I. = 1.00, change G to C); 355 (C.I. = 1.00, change G to A); 394 (C.I. = 0.50, change C to G); and 437 (C.I. = 0.333, change A to G). The exclusion of any of these characters transferred part of the *Herrania* outgroup (cluster 'section *Herrania*') to the ingroup with bootstrap of 100% (results not shown).

The most parsimonious trees derived from the nucleotide sequences corroborated the sectional classification of *Theobroma* and *Herrania*, which had been based on morphology (Fig. 3). The species were separated into three distinct clades. The *Herrania* species grouped consistently in a *Herrania* clade, separated from all *Theobroma* species. The *Herrania* species were further separated according to sections into 'section *Subcymbicalyx*' (bootstrap = 100%, D.I. = 14) and 'section *Herrania*' (bootstrap = 100%, D.I. = 12). The *Theobroma* species were consistently divided into two clades. One contained all species from the sections considered by Cuatrecasas (1964) as more derived (*Oreanthes*, *Rhytidocarpus*, *Telmatocarpus*, and *Theobroma*), while the second cluster was formed by species from sections exhibiting more plesiomorphic features (*Glossopetalum* and *Andropetalum*). The species of section *Oreanthes* (*T. sylvestre* and *T. speciosum*) were consistently grouped (bootstrap = 99%, D.I. = 9), and both were strongly clustered with *T. bicolor* (*Rhytidocarpus*) [bootstrap = 95%, D.I. = 6]. Grouping of these sections with *T. microcarpum* (section *Telmatocarpus*) was inconsistent (bootstrap not significant).

Within the group formed by the most ancestral sections, there was no consistent separation between species restricted to Central America (*T. simiarum*, *T. angustifolium*, and *T. mammosum*) and those occurring in the region east of the Andes (*T. grandiflorum*, *T. obovatum*, and *T. subincanum*) [bootstrap = 99%, D.I. = 8]. The Central American species (*T. angustifolium* and *T. mammosum*) were paraphyletic to the species that occur east of the Andes (bootstrap = 90%, D.I. = 3) (Fig. 3).

Altering the weight of the characters in the parsimony analysis (2:1:1 for each codon base; and 1:2 step-matrix for transition:transversions) resulted in either identical or compatible trees with the consensus trees generated by simple parsimony (equal weight for characters).

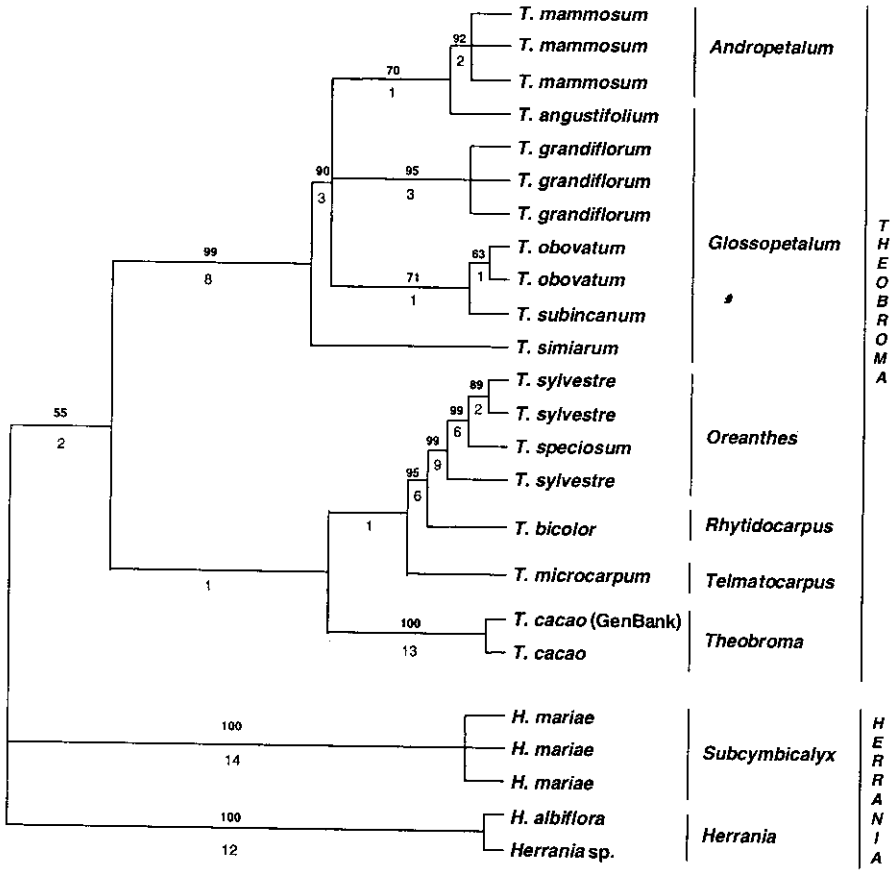

Fig. 3. Strict consensus of 16 most parsimonious trees generated from nucleotide sequence of the trypsin inhibitor sequences from *Theobroma* and *Herrania* species. Length = 273 steps; C.I. = 0.7839; H.I. = 0.2161; R.I. = 0.8628; R.C. = 0.6763. Percent of 1,000 bootstrap replications is given above the branches. Decay index values are shown below the branches. Classification into sections followed Cuatrecasas (1964) for *Theobroma* and Schultes (1958) for *Herrania*

The parsimony analyses of the deduced amino acid sequence resulted in a consensus tree congruent to the one derived from nucleotide sequence, but part of the outgroup *Herrania* (cluster 'section *Subcymbicalyx*') was transferred to the ingroup, at the base of *Theobroma* (bootstrap = 100%). The remaining outgroup (cluster 'section *Herrania*') showed an increase in resolution, linking consistently at the base of the ingroup (bootstrap = 100%) [results not shown].

Discussion

Proteinase inhibitors are present in seeds of many plant species, where they might act as anti-herbivore compounds; as sulfur storage; as regulators of endogenous proteases; or any combination of these functions (Baek et al. 1994, Hammond et al. 1984). The serine-proteinase inhibitors are a well-characterized class of inhibitors, which can be classified into at least 16 families based on the binding mechanism to the target protein and on sequence similarity (Bode and Huber 1992). Among the plant serine-proteinase inhibitors, the soybean Bowman-Birk and the Kunitz trypsin inhibitors are two of the best characterized (Ryan 1990). The 21 kDa peptide from *T. cacao* seeds had been previously characterized as a member of the Kunitz trypsin inhibitor family (Spencer and Hodge 1991, Tai et al. 1991).

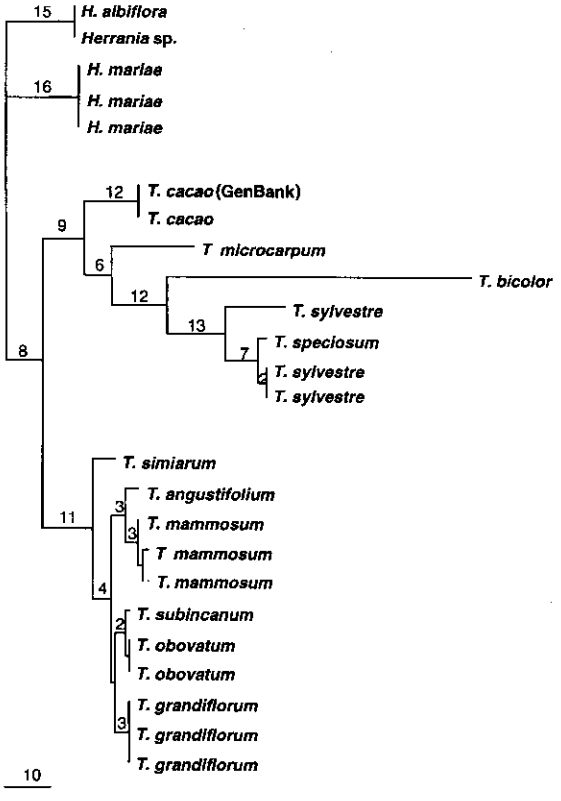

Fig. 4. Phylogram of one of the 16 equally parsimonious trees based on nucleotide sequence of the trypsin inhibitor from species of the genera *Theobroma* and *Herrania*. Branch lengths (optimized by ACCTRAN) are indicated in scale above branches

Multiple alignment of the presumed amino acid sequences of the putative trypsin inhibitor from *Theobroma* and *Herrania* species demonstrated a high similarity between these sequences and the one from *T. cacao* (GenBank X56509.1). The nucleotide sequence of the clone obtained in this study from cacao was identical to the previously characterized cDNA, validating the approach to identify the trypsin inhibitor gene sequences from other species. None of the nucleotide sequences presented introns, as described for the *T. cacao* gene (Spencer and Hodge 1991, Tai et al. 1991) and other plant trypsin inhibitor genes (Ishikawa et al. 1994, Jofuku and Goldberg 1989).

Low-copy nuclear genes exhibit higher rates of sequence divergence than chloroplast and nuclear ribosomal genes, being useful in resolving close interspecific relationships (Sang 2002). To compare the evolutionary rate of the trypsin inhibitor gene for this group of species, the pairwise *p*-distance was contrasted with those obtained for the vicilin gene and the chloroplastic *ndh*F published by Whitlock and Baum (1999). Using a single representative from *Theobroma* (*T. cacao*) and from *Herrania* (*Herrania* sp.), the uncorrected pairwise distance (*p*-distance) was 7.9% for the trypsin inhibitor gene, 5.9% for the vicilin gene, and 0.75% for the *ndh*F. Therefore, the trypsin inhibitor gene has accumulated mutations at a 1.3 faster rate than the vicilin gene and 10.5 times faster than the *ndh*F gene.

Some within-individual and infraspecific sequence diversity was detected in a few clones suggesting that this gene is encoded by a small gene family. According to Dodo et al. (1994), there are between 5 to 7 copies of the trypsin inhibitor gene in *T. cacao* detected by Southern analysis. Similar results have been observed for the *Theobroma* species (unpublished results). The Kunitz trypsin inhibitor genes are encoded by a small gene family in other plant species (Garcia-Omeldo et al. 1987, Jofuku and Goldberg 1989, Ryan 1990).

The consistent correlation between the phylogenetic tree for the trypsin inhibitor nucleotide sequences from *Theobroma* and *Herrania* and morphological data, suggests that the phylogeny of the gene is congruent with the phylogeny of the species. The data for the trypsin inhibitor sequences corroborated the results of the vicilin gene analysis (Whitlock and Baum 1999) and morphology (Cuatrecasas 1964). *Herrania* and *Theobroma* are closely related, but the monophyly of *Theobroma* was not strongly supported (bootstrap = 55%, D.I. = 2). The unbranched trunk; compound palmate leaves; and long petal-lamina, exceeding the length of the petal-hood are the major characters that distinguish *Herrania* from *Theobroma*, representing probable synapomorphies of *Herrania* (Schultes

1958, Cuatrecasas 1964). Cuatrecasas (1964) suggested that there were no obvious morphological synapomorphies for *Theobroma*, a conclusion congruent with the phylogenetic analysis of the trypsin inhibitor gene sequence here presented, and the vicilin gene (Whitlock and Baum 1999). The phylogenetic analysis of the chloroplastic gene *ndh*F confirmed that both genera are closely related, but it was not able to resolve the relationships between the species of *Theobroma* and *Herrania* (Whitlock et al. 2001).

The classification of *Theobroma* and *Herrania* into sections was corroborated in this study. It is noteworthy that the species were divided in both genera into two clades. The *Herrania* clade was sub-divided into the relatively plesiomorphic section *Subcymbicalyx* and the section *Herrania*, with apomorphic characteristics (Figs. 3 and 4). The *Theobroma* clade was sub-divided into a group containing sections *Andropetalum* and *Glossopetalum*, which exhibit many plesiomorphic characteristics, and a second group of the morphologically derived sections *Oreanthes*, *Rhytidocarpus*, *Telmatocarpus*, and *Theobroma*. *Theobroma microcarpum* (*Telmatocarpus*) grouped inconsistently (bootstrap not significant) with the more derived sections, corroborating the ambiguous evolutionary classification of this section according to morphology (Cuatrecasas 1964); seed fatty acid profile (Silva et al. 2001); and purine alkaloid composition (unpublished results). The classification of *Theobroma* into sections based on morphology had also been corroborated by rDNA polymorphism (Figueira et al. 1994) and by fatty acid and sterol composition of seeds (Carpenter et al. 1994). Other studies using chemical and molecular data, including fatty acid composition, total seed protein profile, random amplified polymorphic DNA (RAPD) (Silva et al. 2001) and vicilin gene (Whitlock and Baum 1999) have given support to the established classification of *Theobroma*.

Our data supported the monophyly of all *Theobroma* sections, except for *Glossopetalum*, which was paraphyletic to section *Andropetalum* (*T. mammosum*), in agreement with the results of Whitlock and Baum (1999) based on the vicilin gene. *Theobroma mammosum*, the only member of *Andropetalum*, a section described by Cuatrecasas (1964), is endemic to Costa Rica. Section *Andropetalum* differed from species of *Glossopetalum* by only a few characters (extraordinarily broad and reflexed staminodes; relatively reduced petal-lamina, and gamosepalous calyx). However, *Glossopetalum* is also characterized by other features common to *Andropetalum* such as the presence of large and reflexed staminodes at floral bud stage, and reflexed and erect at anthesis, petal-hood 7-nerved, fruits with rigid pericarp, hard epicarp, and tomentose leaves at abaxial side (Cuatrecasas 1964). The close genetic relationship between *Glossopetalum* and *Andropetalum* may also be inferred by the existence of inter-specific hybrids between *T. angustifolium* (*Glossopetalum*) and *T. mammosum* (*Andropetalum*); and hybrids between *T. simiarum* (*Glossopetalum*) and *T. mammosum* (*Andropetalum*) (Cuatrecasas 1964). The morphological characteristics, the inter-specific hybridization between sections, associated with the results of the phylogenetic analyses of *Theobroma* based on the sequence of the trypsin inhibitor gene here presented, and the vicilin gene (Whitlock and Baum 1999), support the inclusion of section *Andropetalum* within *Glossopetalum*. If *T. mammosum* were considered a member of *Glossopetalum*, there would be a perfect monophyly for all sections of the genus *Theobroma*. The characters utilized by Cuatrecasas (1964) to propose section *Andropetalum*, distinguishing this section from species of *Glossopetalum*, are autapomorphic, and are thus only useful to give identity to *T. mammosum*. Therefore, *Andropetalum* may be considered a synonym of section *Glossopetalum*.

According to Cuatrecasas (1964), the elevation of the Andes separated previously widespread *Theobroma* populations, favoring speciation by isolation. Currently, except for the cultivated species (see introduction), the *Theobroma* species that occur east of the Andes do not occur on the west side.

In addition, east/west vicariant species pairs exist (Cuatrecasas 1964). Only three *Theobroma* species (*T. simiarum*, *T. angustifolium*, and *T. mammosum*), from the most plesiomorphic section (*Glossopetalum*), occur naturally in Central America, while all the other species, including the most derived ones, exhibit a distribution limited to South America. A similar trend occurs in *Herrania*, considered by Cuatrecasas (1964) to be a more ancestral genus in relation to *Theobroma*, with one species occurring in Central America (*H. purpurea*), and the other 16 limited to South America, on both sides of the Andes (Schultes 1958). In our analysis, there was no consistent separation of species from *Glossopetalum*, including *T. mammosum*, of exclusive occurrence in Central America, from those of the region east of the Andes (*T. grandiflorum*, *T. subincanum*, and *T. obovatum*). Thus, it can be hypothesized that an ancestral population once extended from the Amazon region through Central America, which ultimately developed into the two genera, *Herrania* and *Theobroma*. The derived *Theobroma* species, including *T. cacao*, might have evolved only in South America due to geographical restriction imposed by the Andes and the Panama isthmus. Recent molecular evidences have suggested that domesticated cacao might in fact have originated from a few individuals from South America that were carried into Central America by humans (Motamayor et al. 2002).

In this study, the potential use of the low-copy nuclear trypsin inhibitor gene sequences in phylogenetic studies of *Theobroma* and *Herrania* was demonstrated, in agreement with Sang (2002). However, questions concerning the monophyly of *Theobroma* remain, although it was tentatively supported as monophyletic. *Theobroma mammosum*, and consequently section *Andropetalum* may be included into section *Glossopetalum*.

We acknowledge the financial support by 'Conselho Nacional de Desenvolvimento Científico e Tecnológico' (CNPq); 'Fundação de Amparo à Pesquisa do Estado de São Paulo' (FAPESP); International Foundation of Science (IFS), and the permission from Empresa Brasileira de Pesquisas Agropecuária (EMBRAPA); Comissão Executiva do Plano da Lavoura Cacaueira (CEPLAC) to use the plant material. We thank the useful comments from the reviewers.

References

Addison G. O., Tavares R. M. (1951) Observações sobre as espécies do gênero *Theobroma* que ocorrem na amazônia. Bol. Tec. Inst. Agronômico do Norte 25: 3–20.

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

Alverson W. S., Whitlock B. A., Nyffler R., Bayer C., Baum D. A. (1999) Phylogeny of the core Malvales: evidence from *ndh*F sequence data. Amer. J. Bot. 86: 1474–1486.

Baek J. M., Song J. C., Choid Y. D., Kim S. I. (1994) Nucleotide sequence homology of cDNAs encoding soybean Bowman-Birk type proteinase inhibitor and its isoinhibitors. Biosci. Biotech. Bioch. 58: 843–846.

Bayer C., Fay M. F., De Bruijn A. Y., Savolainen V., Morton C. M., Kubitzki K., Alverson W. S., Chase M. W. (1999) Support for an expanded family concept of Malvaceae within a recircumscribed order Malvales: a combined analysis of plastid *atp*B and *rbc*L DNA sequences. Bot. J. Linn. Soc. 129: 267–303.

Bode W., Huber R. (1992) Natural protein proteinase inhibitors and their interactions with proteinases. Eur. J. Biochem. 204: 433–451.

Bremer K. (1988) The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. Evolution 42: 795–803.

Carpenter D. R., Hammerstone J. F., Romanczyk L. J., Aitken W. M. (1994) Lipid composition of *Herrania* and *Theobroma* seeds. J. Am. Oil Chem. Soc. 71: 845–851.

Cuatrecasas J. (1964) Cacao and its allies: a taxonomic revision of the genus *Theobroma*. Contrib. US Natl. Herbarium 35: 379–614.

Dodo H. W., Fritz P. J., Furtek D. B. (1992) A cocoa 21 kilodalton seed protein has trypsin inhibitory activity. Café Cacao Thé 36(4): 279–284.

Dodo H. W., Tai H., Fritz P., Furtek D. (1994) Studies of the protein and gene encoding a 21 Kilodalton trypsin inhibitor from cocoa seeds. In: Proceedings 11th International Cocoa Research Conference - Yamassoukro, Ivory Coast, July 18-24th 1993. Cocoa Producers' Alliance, Lagos, Nigeria, pp. 1–10.

Doyle J. J., Doyle J. L. (1990) Isolation of plant DNA from fresh tissue. Focus 12: 13–15.

Figueira A., Janick J., Levy M., Goldsbrough P. B. (1994) Reexamining the classification of *Theobroma cacao* L. using molecular markers. J. Amer. Soc. Hortic. Sci. 119: 1073–1082.

Garcia-Olmedo F., Salcedo G., Sanchez-Monge R., Gomez L., Royo J., Carbonero P. (1987) Plant proteinaceous inhibitors of proteinases and alpha-amylases. Oxford Surv. Plant Mol. Cell Biol. 4: 275–334.

Hammond R. W., Foard D. E., Larkins B. A. (1984) Molecular cloning and analysis of a gene coding for the Bowman-Birk protease inhibitor in soybean. J. Biol. Chem. 259: 9883–9890.

Ishikawa A., Ohta S., Matsuoka K., Hattori T., Nakamura K. (1994) A family of potato genes that encode Kunitz-type proteinase inhibitors - structural comparisons and differential expression. Plant Cell Physiol. 35: 303–312.

Jofuku K. D., Goldberg R. B. (1989) Kunitz trypsin inhibitor genes are differentially expressed during the soybean life cycle and in transformed tobacco plants. Plant Cell 1: 1079–1093.

Judd W. S., Manchester S. R. (1997) Circumscription of Malvaceae (Malvales) as determined by a preliminary cladistic analysis of morphological, anatomical, palynological, and chemical characters. Brittonia 49: 384–405.

McHenry L., Fritz P. J. (1992) Comparison of the structure and nucleotide sequence of vicilin genes of cocoa and cotton raise questions about vicilin evolution. Plant Mol. Biol. 18: 1173-1176.

Motamayor J. C., Risterucci A. M., Lopez P. A., Ortiz C. F., Moreno A., Lanaud C. (2002) Cacao domestication I: the origin of the cacao cultivated by the Mayas. Heredity 89: 380–386.

Purseglove J. W. (1968) Tropical Crops-Dicotyledons, vol. 1. Longman Green London, London.

Ryan C. A. (1990) Protease inhibitors in plants: Genes for improving defenses against insects and pathogens. Ann. Rev. Phytopathol. 28: 425–449.

Salisbury B. A. (2000) Strongest evidence revisited. Cladistics 16: 394–402.

Sang T. (2002) Utility of low-copy nuclear gene sequences in plant phylogenetics. Crit. Rev. Biochem. Mol. Biol. 37 (3): 121–147.

Schultes A. (1958) Synopsis of the genus *Herrania*. J. Arnold Arboretum 39: 216–278.

Silva C. R. S., Figueira A. V. O., Souza E. C. A. S. (2001) Diversidade no gênero *Theobroma*, In: Dias L.A.S. (ed.) Melhoramento genético do cacaueiro, FUNAPE - UFG, Goiânia, pp. 49–80.

Spencer M. E., Hodge R. (1991) Cloning and sequencing of the cDNA encoding the major albumin of *Theobroma cacao*. Planta 183: 528–35.

Swofford D. L. (1998) PAUP* - Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Tai H., McHenry L., Fritz P. J., Furtek D. B. (1991) Nucleic acid sequence of a 21 kDa cocoa seed protein with homology to the soybean trypsin inhibitor (Kunitz) family of protease inhibitors. Plant Mol. Biol. 16: 913–915.

Thompson J. D., Gibson T. J., Plewniak F., Jeanmougin F., Higgins D. G. (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 24: 4876–4882.

Voigt J., Biehl B. (1993) The major seed proteins of *Theobroma cacao* L. Food Chem. 47: 145–151.

Whitlock B. A., Baum D. A. (1999) Phylogenetic relationships of *Theobroma* and *Herrania* (Sterculiaceae) based on sequences of the nuclear gene vicilin. Syst. Bot. 24: 128–138.

Whitlock B. A., Bayer C., Baum D. A. (2001) Phylogenetic relationships and floral evolution of the Byttnerioideae ("Sterculiaceae" or Malvaceae s.l.) based on sequences of the chloroplast gene, *ndh*F. Syst. Bot. 26: 420–437.

Address of the authors: Carlos Rogério Sousa Silva (e-mail: roger@cena.usp.br), Antonio Figueira (e-mail: figueira@cena.usp.br), Laboratório de Melhoramento de Plantas, Centro de Energia Nuclear na Agriculture, Universidade de São Paulo, Av. Centenário, 303, CP 96, Piracicaba, São Paulo, 13400-970, Brazil.