Genetica (2009) 135:123–135
DOI 10.1007/s10709-008-9257-7

Molecular evolution of paclitaxel biosynthetic genes *TS* and *DBAT* of *Taxus* species

Da Cheng Hao · Ling Yang · Beili Huang

Received: 14 October 2007 / Accepted: 27 February 2008 / Published online: 8 March 2008

Abstract Evolutionary patterns of sequence divergence were analyzed in genes from the conifer genus *Taxus* (yew), encoding paclitaxel biosynthetic enzymes taxadiene synthase (TS) and 10-deacetylbaccatin III-10β-O-acetyltransferase (DBAT). N-terminal fragments of *TS*, full-length *DBAT* and internal transcribed spacer (ITS) were amplified from 15 closely related *Taxus* species and sequenced. Premature stop codons were not found in *TS* and *DBAT* sequences. Codon usage bias was not found, suggesting that synonymous mutations are selectively neutral. *TS* and *DBAT* gene trees are not consistent with the ITS tree, where species formed monophyletic clades. In fact, for both genes, alleles were sometimes shared across species and parallel amino acid substitutions were identified. While both *TS* and *DBAT* are, overall, under purifying selection, we identified a number of amino acids of *TS* under positive selection based on inference using maximum likelihood models. Positively selected amino acids in the N-terminal region of TS suggest that this region might be more important for enzyme function than previously thought. Moreover, we identify lineages with significantly elevated rates of amino acid substitution using a genetic algorithm. These findings demonstrate that the pattern of adaptive paclitaxel biosynthetic enzyme evolution can be documented between closely related *Taxus* species, where species-specific taxane metabolism has evolved recently.

Keywords Adaptive evolution · 10-Deacetylbaccatin III-10β-O-acetyltransferase · Paclitaxel · Positive selection · Taxadiene synthase · *Taxus*

D. C. Hao
Laboratory of Pharmaceutical Resource Discovery, Dalian Institute of Chemical Physics, Dalian 116023, China
e-mail: haodc@126.com

L. Yang (✉)
Laboratory of Pharmaceutical Resource Discovery, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian, China
e-mail: yling@dicp.ac.cn

B. Huang
Lushan Botanical Garden, Chinese Academy of Sciences, Jiangxi, China

Introduction

Plants synthesize an enormous number of secondary compounds that provide an increasingly exploited reservoir for the generation of pharmaceutically active agents (Hartmann et al. 2005), and many more await discovery. In the conifer genus *Taxus*, paclitaxel (Taxol), a well-known anti-cancer agent, and related taxane compounds are major components in the mixture of secondary metabolites, which play an important ecological role in plant defense. There is a large variation in taxane content among the different species and cultivars (van Rozendaal et al. 2000). Over the past few years, major advances have been made in the identification of genes responsible for paclitaxel biosynthesis, a process requiring an estimated dozen enzymatic reactions involving the construction of the tetracyclic skeleton and the addition of the various oxygen and acyl functional groupings. Among the intermediate steps, the cyclization of geranylgeranyl diphosphate to taxadiene is catalyzed by taxadiene synthase (TS; Koepp et al. 1995), and the acetylation of 10-deacetyl baccatin III to baccatin III is catalyzed by 10-deacetyl baccatin III-10-O-acetyltransferase (DBAT).

The cDNA sequence of *Taxus brevifolia TS* specifies an open reading frame of 2586 nucleotides, 13 exons, and the

deduced full-length preprotein (862 residues, 98.3 kDa) includes an N-terminal plastid targeting sequence, a conifer diterpene internal sequence domain (CDIS) encoded by exons 2–4, an internal glycosyl hydrolase-like domain, an active site domain encoded by exons 10–13, and the typical terpene synthase DDXXD divalent metal ion-substrate complex binding motif (Trapp and Croteau 2001). Deletion of up to 79 N-terminal residues yielded functional protein; however, deletion of 93 or more amino acids resulted in complete elimination of activity, implying a structural or catalytic role for the amino terminus (Williams et al. 2000). Comparison of the translated *TS* sequence to other terpene synthase sequences shows significant homology to abietadiene synthase (46% identity, 67% similarity) from grand fir (Wildung and Croteau 1996).

The full-length cDNA of *Taxus cuspidata DBAT* has an open reading frame of 1320 base pairs corresponding to a deduced protein of 440 residues with a calculated molecular weight of 49,052, consistent with the size of the operationally soluble, monomeric native acetyltransferase demonstrated in *Taxus* cell extracts (Walker and Croteau 2000). DBATs contain a highly conserved HXXXDG sequence motif found in other transacylases. The recombinant DBAT has a pH optimum at 7.5, Km values of 10 and 8 μM for 10-deacetylbaccatin III and acetyl coenzyme A, respectively, and is seemingly regiospecific towards the 10β-hydroxyl group of the taxane ring (Walker and Croteau 2000).

Positive, diversifying selection is an important evolutionary force that accelerates divergence between homologous proteins (Swanson et al. 2001). Among the proteins identified to be under positive selection are immune-response and defense-related genes (Bishop 2005; Nielsen et al. 2005), and toxin protein genes (Liu et al. 2005). Since paclitaxel biosynthetic enzymes catalyze the formation of an important defense molecule paclitaxel and other related taxanes, it is reasonable to expect that most amino acid residues are highly conserved. However, whether adaptive evolution affects a few sites of some enzymes is unknown. As far as we know, there is no study addressing patterns of evolution of the paclitaxel biosynthetic enzymes within the genus *Taxus*, although knowledge about codons that are under positive selection and purifying selection is important for studies of plant secondary metabolism and phylogenetics and could facilitate the development of more broadly applicable enzymes for biotransformation of taxanes. The goal of this study was to determine patterns of evolution of *TS* and *DBAT* in *Taxus*. We identify differences in tree topologies between these biosynthetic enzymes and a commonly used phylogenetic marker, nuclear internal transcribed spacer (ITS). We document positive selection acting on *TS* but not on *DBAT* and determine the identity of amino acid sites under selection and parallel substitution. Finally, we identify lineages with significantly elevated rates of amino acid substitution of *TS*. Together, these results suggest that positive selection is driving divergence of *TS* in closely related *Taxus* species and allow us to nominate candidate amino acid sites that may contribute to the differential taxane metabolism between sister taxa.

Materials and methods

Sampling, amplification, and sequencing

Species, geographic origin of the sequenced material, their voucher numbers, and GenBank accession numbers of the sequences generated in this study, as well as those retrieved from GenBank, are given in Table 1; 12 *TS*, 14 *DBAT*, and 16 ITS sequences were newly generated for this study.

Samples of *Taxus* were identified by taxonomic characters in Spjut (2007a; http://www.worldbotanical.com/Nomenclature.htm#nomenclature); however, most samples of species and varieties are from geographical areas where there is not likely to be a problem with sympatric taxa as determined from data in Spjut (2007a, b). These include the North American *T. brevifolia*, *T. globosa* Schltdl. var. *globosa*, *T. globosa* var. *floridana* (Nutt. ex Chapm.) Spjut, and *T. canadensis*, the Northwest Himalayan *T. contorta* from Jilong, Tibet, the East Himalayan *T. wallichiana* from ChaYu, Tibet, and *T. chinensis* from Hubei. *Taxus wallichiana* var. *yunnanensis* was distinguished from var. *wallichiana* by the relatively thinner leaves with revolute margins; its identifications were further confirmed by anatomical character features in Spjut (2007b). We also included cultivars *T.* × *hunnewelliana* Rehder, *T.* × *media* Rehder and species from Eurasia that are cultivated, *T. baccata*, *T. recurvata*, and *T. cuspidata* Siebold & Zucc., the identifications of which are all based on the name at the source where the plant is grown. The origin for the sample of *T. sumatrana* (Miq.) de Laub. is unknown.

Genomic DNA was extracted by using Universal Genomic DNA Extraction kit (Takara, Dalian, China), following the manufacturer's protocol. A 0.9% agarose gel was run to assess the presence and integrity of the DNA. Quantification was done spectrophotometrically and the concentration of DNA ranged from 50–77 ng per μl.

A 50 μl PCR reaction mix consisted of 5 μl of 10 × reaction buffer, 4 μl each of 2.5 mM dNTPs stock, 2.5 μl of 10 μM forward and reverse primers (synthesized by Takara, Dalian, China), 0.5 μl bovine serum albumin (10 mg/ml), and 1.5 units of Ex Taq polymerase (Takara, Dalian, China). The exons 1–4 of *TS* gene (Fig. 1) were amplified using 5′-atggctcagctctcatttaatgc (forward) and 5′-cgcagccgccgaatttgtcca (reverse). The exons 5–9 of *TS*

Table 1 Samples of 15 *Taxus* species

Taxon	Origin	Voucher no.	*TS*	*DBAT*	ITS
T. × *hunnewelliana* Rehder	Waterloo, Canada	WC001 (WAT)	**EU107120**	**EU107132**	**EF660579**
	Vancouver, Canada	UBC200707 (UBC)	**EU107121**	**EU107133**	ND
T. chinensis Pilger	ShenNongJia, HuBei	SNJ001	AY007207	**EU107135**	**EF660597**
	YunXi, HuBei	YX001	**EU107126**	**EU107142**	ND
T. × *media* Rehder	Dalian, China	DICP001	AY461450	EF028093	**EF660598**
T. cuspidata Siebold & Zuccarini	Ji'An, JiLin, China	JA001, http://www.worldbotanical.com/Taxus%20images/Taxus_umbraculifera-JA001.jpg	DQ305407	AF193765	**EF660602**
T. cuspidata var. *nana* Rehder	Japan	DD001	**EU107124**	–	**EF660576**
T. wallichiana Zuccarini var. *yunnanensis* (W. C. Cheng & L. K. Fu) C. T. Kuan	ChaYu, Tibet, China	CY001	**EU107129**	**EU107136**	**EF660568**
T. recurvata Spjut	Oxford, UK	Stevenson, 0000381 (OXF) http://www.worldbotanical.com/Taxus%20images/Taxus_recurvata0000381.jpg	AY424738	AF456342	**EF660599**
T. baccata Linnaeus	Montreal, Canada	Bailleul, 1586–1978 (MTJB) http://www.worldbotanical.com/Taxus%20images/Taxus-baccata-1586-1978.jpg	**EU107123**	**EU107141**	ND
T. canadensis Marshall	Montreal, Canada	Bailleul, 1960–2000 (MTJB) http://www.worldbotanical.com/Taxus%20images/Taxus_canadensis-Bailleul.jpg	AY364470	**EU107134**	**EF660601**
T. mairei (Lemée & H. Léveillé) S. Y. Hu ex T. S. Liu	LiShui, ZheJiang, China	LS001 http://www.worldbotanical.com/Taxus%20images/Taxus_chinensis-LS001.jpg	AY931015	AY365031	**EF660596**
	JiangXi, China	JX001 http://www.worldbotanical.com/Taxus%20images/Taxus_mairei-1.jpg	**EU107125**	**EU107140**	ND
T. obscura Spjut	NanPing, FuJian, China	NP001 http://www.worldbotanical.com/Taxus%20images/Taxus_obscura-NP001.jpg	**EU107127**	**EU107139**	ND

Table 1 continued

Taxon	Origin	Voucher no.	*TS*	*DBAT*	ITS
T. contorta Grif.	JiLong, Tibet, China	JL001	**EU107122**	**EU107138**	**EF660582**
		JL002	**EU107128**	**EU107137**	ND
T. wallichiana Zuccarini	ChaYu, Tibet, China	CYW001	**EU107130**	–	**EF660573**
T. sumatrana (Miquel) de Laubenfels	Unknown	Determann, ABG20051056 (ATLAN)	**EU107131**	**EU107144**	**EF660572**
T. globosa Schlectendahl	Mexico	Determann, ABG19971263 (ATLAN)	–	**EU107145**	**EF660570**
T. brevifolia Nuttall	Vancouver, Canada	La Fountaine, UBC20070701 (UBC)	U48796	**EU107143**	**EF660598**
T. floridana Nuttall ex Chapman	Gainesville, USA	Spjut, 12172	–	–	**EF660603**
Outgroup					
Abies grandis (Douglas ex D. Don) Lindley, *abietadiene synthase*			U50768		
Austrotaxus spicata Compton	New Caledonia	Determann, ABG20060714 (ATLAN)	–	–	**EF660569**

Accession numbers of new sequences are given in boldface. All unmarked vouchers are deposited in Herbarium, Lushan Botanical Garden, Chinese Academy of Sciences, Jiangxi, China (LUS). A dash indicates missing data. ND, not done

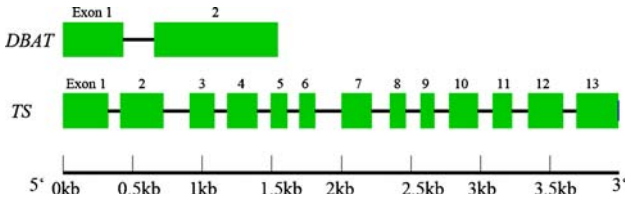

Fig. 1 Exon/intron structure of *TS* and *DBAT* genes in *Taxus* drawn to scale. Boxes indicate exons; areas amplified for interspecific analysis are exons 1 and 2 of *DBAT* and exons 1–9 of *TS*

gene were amplified using 5′-tggacaaattcggcggctgcg (forward) and 5′-cttgttggaagcttcaactcctc (reverse). This segment was amplified from *T. wallichiana* var. *yunnanensis* and *T.* × *hunnewelliana* WC001 and the corresponding sequences of other seven species (*T.* × *media*, *T. baccata*, *T. mairei*, *T. cuspidata*, *T. chinensis*, *T. brevifolia*, and *T. canadensis*) were retrieved from GenBank (Table 1). The *DBAT* gene (Fig. 1) was amplified using 5′-atggcaggctcaacagaatttg (forward) and 5′-tcaaggtttagttacatatttgtttg (reverse). Approximately 50 ng of genomic DNA was used as a template for the reaction. The reaction mixture was placed in a Takara PCR Thermal Cycler Dice (Takara, Japan). PCR conditions: 94°C for 5 min; 38 cycles at 94°C for 30 s, anneal for 45 s, and 72°C for 1 min 40 s; and finally, 72°C for 7 min. The annealing temperatures were 53°C (*TS*) or 49°C (*DBAT*). ITS was amplified using primers and PCR conditions described previously (Kress et al. 2005). The PCR products were purified by Agarose Gel DNA Purification Kit (Takara).

All PCR products were subcloned into a TA cloning vector pMD19-T (Takara). The plasmids were purified for sequencing. ABI Prism, BigDye Terminator, and cycle Sequencing Ready Reaction Kit (Applied Biosystems, Foster City, CA) were used for sequencing reaction with RV-M and M13-47 primers. The longer clones were sequenced by using oligonucleotides synthesized according to the sequences obtained by RV-M or M13-47 primers. The sequences were detected using an ABI Prism 377 Genetic Analyzer (Applied Biosystems).

Phylogenetic analysis

Sequence alignment was performed with CLUSTAL W and default settings. The aligned *TS* exons 1–4, *TS* exons 1–9, *DBAT*, and ITS matrices comprised 1030, 1689, 1320, and 1246 positions, respectively. Each separate DNA region, as well as all combined data, was analyzed with Modeltest 3.8 (Posada 2006) to find the best model of evolution for the data. Employing the Akaike information criterion (AIC), the model with the lowest AIC score was chosen. Maximum likelihood (ML) and Maximum parsimony (MP) analyses were performed on the separate

molecular partitions and on the combined data. ML analysis and bootstrapping were performed using GARLI 0.951 (Zwickl 2006). GARLI searches relied on the GTR + G, HKY + G, and GTR + I models, which ModelTest selected as the best fitting models for unpartitioned *TS* exons 1–4, *DBAT*, and ITS data, respectively. MP analysis was performed using PAUP* 4.0b10 (Swofford 2002). Heuristic searches were performed using tree bisection–reconnection (TBR) branch-swapping and 10 random sequence addition replicates. All sites were equally weighted and gaps were treated as missing characters. Strong support for individual nodes is defined as nodes with Bayesian posterior probabilities (PP) ≥ 0.95 or non-parametric bootstrap (BP) ≥ 80. Strongly supported conflicting relationships were recovered from *TS* and *DBAT* datasets, so they were not combined for phylogenetic analyses. The data sets were also analyzed with MrBayes 3.1.2 (Ronquist and Huelsenbeck 2003). The analyses of *TS* exons 1–4, *DBAT*, and ITS data utilized one (with outgroup *Abies grandis*, abietadiene synthase gene), three (partitioned by codon position), and one (with outgroup *Austrotaxus spicata*) model partitions, respectively. Two independent runs with one cold and three heated Markov chains each per analysis were performed simultaneously until the average standard deviation of split frequencies between the two runs dropped below 0.01. Analyses were run twice to check for consistency of results. We ran two simultaneous runs for 1.8×10^6 generations and sampled trees every 500 generations. Topology and branch-length information were summarized in 50% majority rule consensus trees; samples obtained before stationarity of –ln likelihoods against generations had been reached were discarded as burn-in. For *TS*, the closely related abietadiene synthase of *Abies grandis* was used as the reference for the rooted tree reconstruction. The ITS sequence of *Austrotaxus spicata* was used as the reference, as *A. spicata* is basal to the genus *Taxus* and is a species within the family Taxaceae (Cheng et al. 2000).

Detection of positive selection and purifying selection, data analysis

We tested for evidence of positive selection by comparing the nonsynonymous substitution rate (d_N) to the synonymous substitution rate (d_S). If a gene is evolving neutrally, $\omega = d_N/d_S$ is expected to equal one, whereas ω greater than one is considered strong evidence that a gene experiences positive selection. We used several ML approaches to test for evidence of positive selection on these taxol biosynthetic enzymes. The first approach, developed by Yang et al. (hereafter referred to as Yang models), involves comparisons of a neutral codon substitution model with ω constrained to be ≤ 1 to a selection model where a class of sites has $\omega > 1$ (Yang et al. 2000). As neutral models are nested within the corresponding selection models, a likelihood ratio test (LRT) can be used to compare them. The test statistic $-2\Delta\ln L$ ($\Delta\ln L$ = the difference in log likelihoods of the 2 models) follows a x^2 distribution with degrees of freedom (df) equal to the difference in number of parameters between models. In the specific models implemented, ω varies between codons as a beta distribution (neutral: M7, M8a; selection: M8). We implemented models M7, M8a, and M8 with the codeml program in PAML4 (Yang 2007). Because Yang models are based on theoretical assumptions and ignore the empirical observation that distinct amino acids differ in their replacement rates, we also implemented MEC (Mechanistic Empirical Combination) model (Doron-Faigenboim and Pupko 2007) that takes into account not only the transition–transversion bias and the nonsynonymous/synonymous ratio, but also the different amino acid replacement probabilities as specified in empirical amino acid matrices. Because the LRT is applicable only when two models are nested and thus is not suitable for comparing MEC and M8a models, the second-order Akaike information criterion (AICc) was used for comparisons (Doron-Faigenboim and Pupko 2007). Those sites that are most likely to be in the positive selection class ($\omega > 1$) are identified as likely targets of selection.

Although the Yang models allow for variation in the nonsynonymous substitution rate, the synonymous rate is fixed across the sequence. Recently, several methods for detecting positive selection that allow for variation in synonymous rate have been proposed. These methods are new implementations of the 3 general classes of previous models, counting methods, fixed effects methods, and random effects methods. Counting methods map changes onto the phylogeny to estimate ω on a site-by-site basis. Kosakovsky Pond and Frost (2005a) propose a version called the single-likelihood ancestor counting (SLAC) method, which calculates the number of nonsynonymous and synonymous substitutions that have occurred at each site using ML reconstructions of ancestral sequences. Kosakovsky Pond and Frost additionally introduce a version of a fixed effect approach, which estimates ω on a site-by-site basis. Their fixed effect likelihood (FEL) method uses ML estimation and treats shared parameters (branch lengths, tree topology, and nucleotide substitution rates) as fixed. The random effects likelihood (REL) method is similar to the Yang model M3; however, both nonsynonymous and synonymous rates vary as gamma distributions with 3 rate classes (Kosakovsky Pond and Frost 2005a). The SLAC, REL, and FEL methods were implemented using the web interface DATAMONKEY (Kosakovsky Pond and Frost 2005b).

HYPHY models that allow d_N/d_S to vary among lineages were used to investigate whether selective pressure on *TS* and *DBAT* genes varies among lineages. The genetic algorithm in HYPHY assigns four classes of d_N/d_S to lineages in a search for "the best model" of lineage-specific evolution (Kosakovsky Pond and Frost 2005c), i.e., $d_N/d_S = 10000$, 1.681, 0.464, and 0, respectively (Fig. 4). This approach can identify lineages under positive selection without an *a priori* hypothesis for lineage-specific evolution.

Parallel amino acid substitutions, codon usage bias

Parallel and convergent evolution refers to independent acquisitions of the same character state on more than one occasion during evolution. The distinction between parallelism and convergence is that the former refers to the situation in which the ancestral states were identical among independent lineages, whereas the latter requires different ancestral states (Zhang 2003). In order to identify parallel amino acid substitutions, ML reconstructions of ancestral sequences and individual mutation events were performed by PAML4 (baseml and pamp). The marginal reconstruction approach (Yang et al. 1995) compares the probabilities of different character assignments to an interior node at a site and select the character that has the highest PP. Number of independent changes, Grantham's distance (Grantham 1974) between starting and ending amino acid, and the possible alternative amino acid substitutions were determined for the identified parallel amino acid substitutions.

Variation in the rate of synonymous substitution among genes may be related to codon use (Sharp 1991). Therefore, several parameters related to codon usage bias for each gene region, such as the codon bias index (CBI; Morton 1993), G + C content at second and third positions as well as overall, and the effective number of codons (ENC; Wright 1990) were estimated using DnaSP version 4.10.4 (Rozas et al. 2003).

Results

Phylogenetic reconstruction

Species formed monophyletic clades on the ITS tree, whose topology (Fig. 2a) is compared to the published ITS tree based on 10 *Taxus* species (Li et al. 2001). The new findings of the present study are: (1) *T. contorta* was basal to the other species in the genus. (2) *T. wallichiana* var. *yunnanensis* and *T. wallichiana* formed a well-supported group that was sister to the group formed by *T. chinensis*, *T. sumatrana*, and *T. mairei*. Li et al. constructed an MP tree and found that three North American species form a well-supported clade, which was also present in our ITS tree (Fig. 2a). However, a few other clades of Li et al.'s ITS tree were only weakly supported. Among them, the clade consisting of *T. cuspidata* and two hybrids was also recovered in our ITS tree, with high PP support. For the grouping of *T. mairei* and *T. chinensis*, there is no conflict between Li et al.'s ITS tree and ours. In Li et al.'s *Taxus* study, the controversial *T. sumatrana*, *T. wallichiana* var. *yunnanensis* and *T. wallichiana* were not included. Due to limited sampling and the MP method used in Li et al.'s study, their results might be less reliable.

Similar to the ITS phylogeny, the *TS* and *DBAT* gene trees generated by different methods did not differ significantly in topology (Fig. 2b, c). On the *TS* tree, the outgroup and three *Taxus* clades formed a polytomy; on the unrooted *DBAT* tree, the polytomy was also observed. Gene trees of *TS* and *DBAT* were not consistent with each other or with the ITS tree. The topology of these gene trees may reflect, (1) cases where the same amino acid substitution occurred independently in more than one lineage, (2) cases of the retention of plesiomorphic characters, and (3) the possibility of incomplete lineage sorting. To minimize the effects of selection, two datasets, one based upon the complete coding sequences and another based upon only third codon positions, were used to generate phylogenetic trees. The *DBAT* tree based on the third codon position was consistent with that based on the complete coding sequence. In contrast, the *TS* tree based on the third codon position was quite different from that based on the complete coding sequence (data not shown), suggesting the strong effect of selection.

TS experienced numerous amino acid substitutions during the evolution of the *Taxus* genus. Sixty-six of 343 amino acid sites (19.2%) in exons 1–4 of TS (Fig. 3a) were variable, with six sites, 147, 193, 276, 295, 298, and 332 having three or four substitutions. For comparison, only 9.8% (43/440) of sites in DBAT were variable and four sites, 216, 219, 352, and 422 having three substitutions (Fig. 3b). However, overall estimates of ω for *TS* (0.468) and *DBAT* (0.349) were less than one, indicating that if these genes experienced positive selection, selection acted on a subset of amino acid sites.

Amino acid sites under selection

Results from all five ML approaches for detecting selection indicated that a proportion of amino acid sites of *TS* have evolved adaptively (Table 2). Model MEC was best-fitting, as the log likelihood value was highest (−2247.64). The LRTs comparing Yang selection model M8 with neutral models (M7 and M8a) were significant (Table 3). Compared to M8a, MEC model had much higher log-likelihood

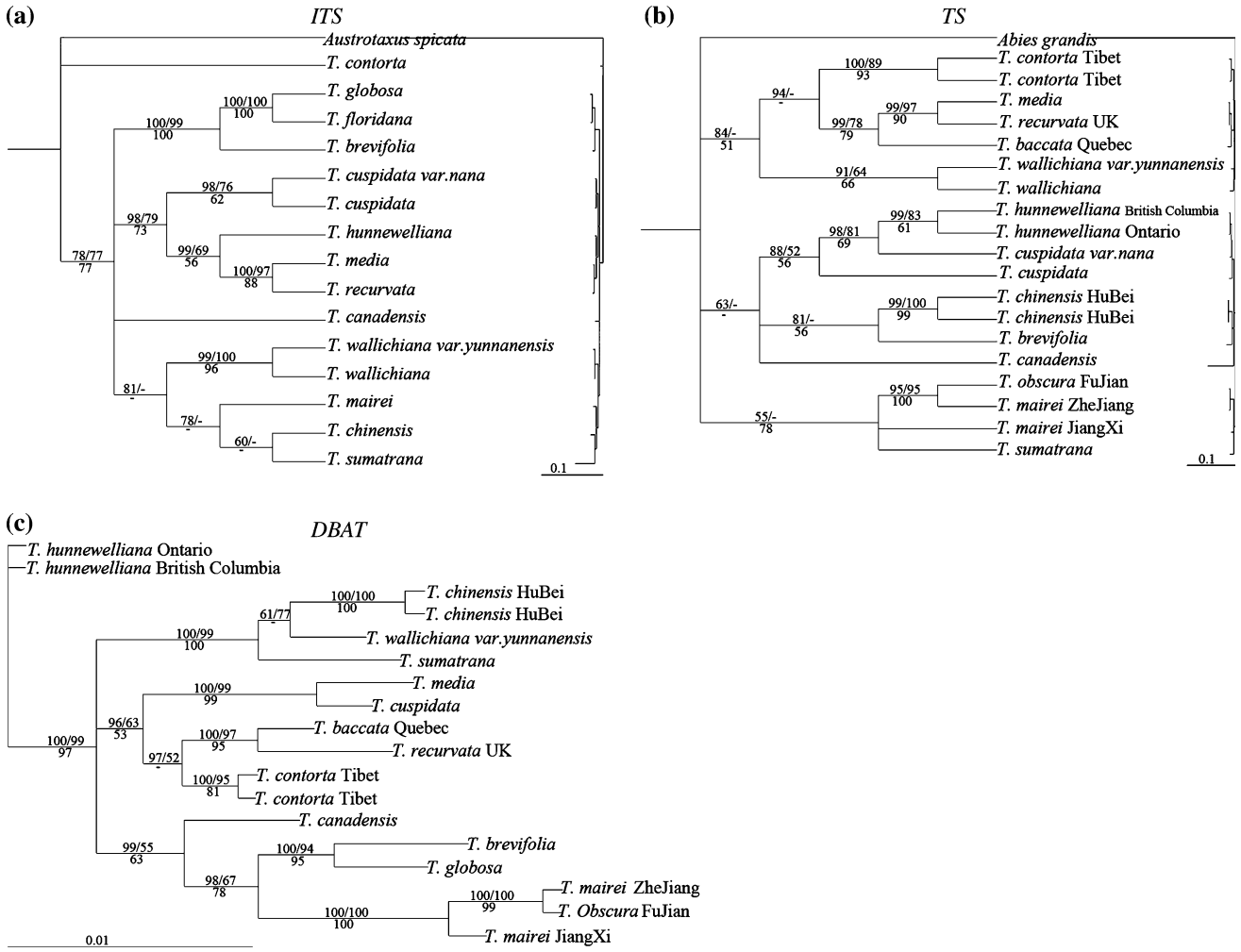

Fig. 2 (**a**) Bayesian 50% majority rule consensus tree (5,600 trees sampled; burn-in = 1,400 trees) inferred from the ITS alignment under the GTR + I model. Bayesian posterior probabilities (PPs, %) are given above branches, before slash (/). ML bootstrap proportions (BPs, %) calculated under the GTR + I model are given above branches, after slash (—, clade not included in the tree). MP BPs are shown below branches. Branch lengths (shown on the right; scale bar, expected number of substitutions per site) are proportional to the mean of the PPs of the branch lengths of the sampled trees. (**b**) Bayesian 50% majority rule consensus tree (3,600 trees sampled; burn-in = 900 trees) inferred from the *TS* (exons 1–4) alignment under the GTR + G model. Bayesian PPs and ML BPs are given above branches (Bayesian/ML). MP BPs are shown below branches. Branch lengths (shown on the right; scale bar, expected number of substitutions per site) are proportional to the mean of the PPs of the branch lengths of the sampled trees. (**c**) Bayesian 50% majority rule consensus tree (3,600 trees sampled ; burn-in = 900 trees) inferred from the *DBAT* alignment under the partitioned model. Bayesian PPs are given above branches, before slash (/). ML BPs calculated under the HKY + G model are given above branches, after slash. MP BPs are shown below branches. Branch lengths (scale bar, expected number of substitutions per site) are proportional to the mean of the PPs of the branch lengths of the sampled trees

value and much lower AICc score. The M8 model identified sites 147, 193, 298, and 332 within exons 1–4 as likely targets of positive selection (Table 2), which were also identified by MEC model. Parameter estimates indicate that the positively selected class has a mean $\omega = 1.646$. The above four sites were reproducibly identified as targets of positive selection when nine *TS* exons 1–9 sequences were analyzed with the above models. The positively selected class has a mean $\omega = 1.990$. For *DBAT*, model M8 was best-fitting (log likelihood value −2514.47). LRT comparing M8 with M8a was not significant (Table 3), and compared to M8a, MEC model had lower log-likelihood value and higher AICc score. Only LRT comparing M8 with M7 was significant ($P < 0.05$). Correspondingly, a few sites identified by the M8 model as likely targets of positive selection were not confirmed by MEC model.

The SLAC method did not identify any sites in TS or DBAT with evidence of positive selection significant at the $P < 0.10$ level; however, site 193 of TS had a $P = 0.22$ of positive selection. Lack of significance at the 0.10 level is not surprising, as counting methods have low power with sequences of low divergence (overall mean distance

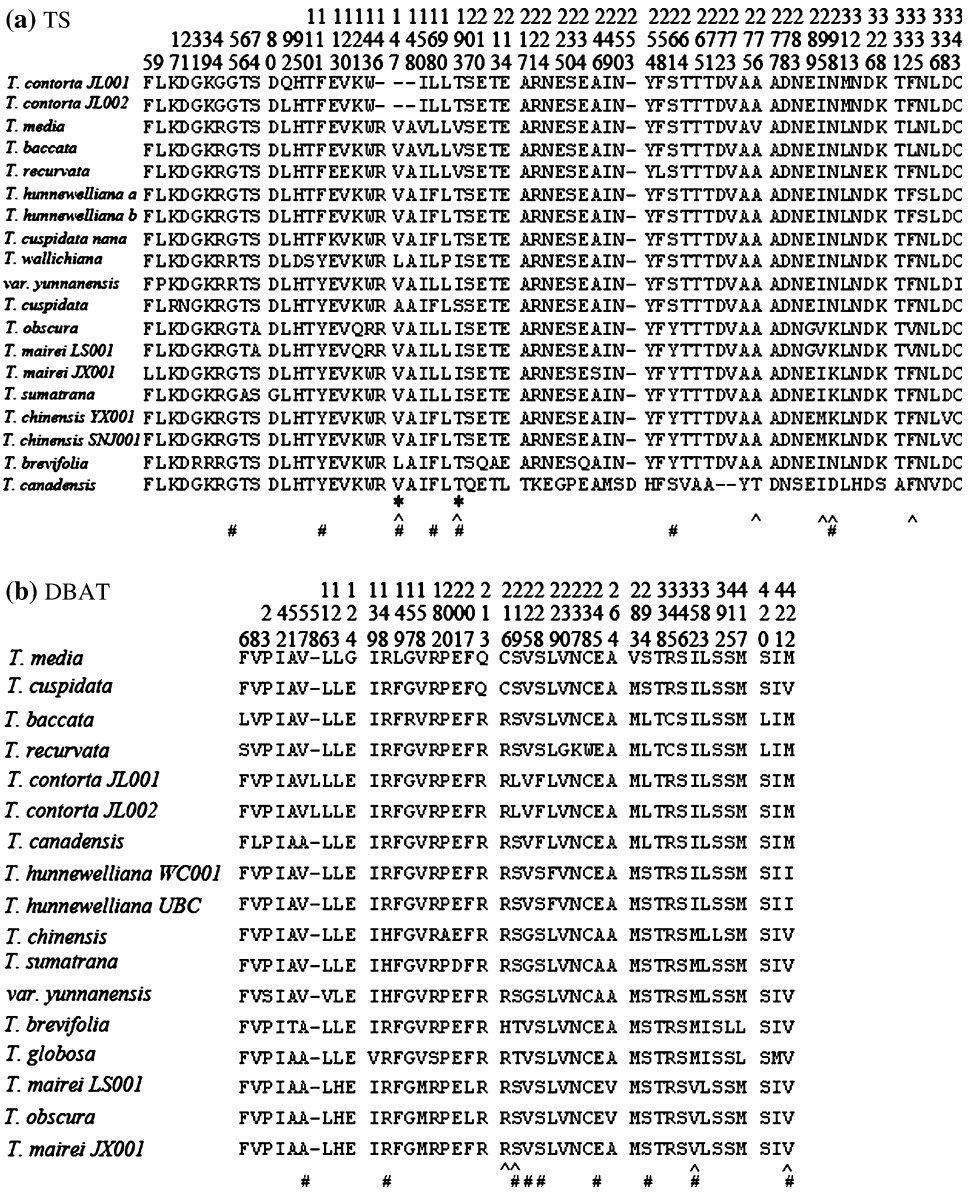

Fig. 3 Amino acid sequence alignment of (**a**) variable amino acid sites in N-terminal region (aa 1–343, encoded by exons 1–4) of TS and (**b**) variable amino acid sites in full length (440 aa) DBAT. Asterisks indicate amino acid sites that have been identified being under positive selection in *Taxus*. Amino acids that have been substituted independently in more than one *Taxus* species are indicated with "#". Sites with at least three substitutions are indicated with arrowheads

calculated by MEGA4, *TS*: 0.01974 ± 0.00222; *DBAT*: 0.01454 ± 0.00175), and analyses of simulated data sets of similar size indicate that *P* values for the SLAC and FEL methods <0.20 have a true Type I error rate of <5% (Kosakovsky Pond and Frost 2005a). Using the FEL method, sites 147 and 193 of TS were significant at the $P < 0.05$ level. The REL method also identified sites 147 and 193 of TS as positively selected. For DBAT, REL identified five sites (219, 228, 294, 352, 422) with PP > 0.5 of positive selection, which was also identified by the M8 model. However, the LRT of M8 vs. M8a and the high AICc score of the MEC model did not justify the positive selection. In addition, the SLAC, FEL, M8, and MEC models identified sites 280I and 361G of DBAT as the negatively selected sites; SLAC, FEL, REL, M8, and MEC models identified sites 173Q, 291L, and 296D of TS as the negatively selected sites.

In summary, all five ML approaches identified site 193 of TS as a likely target of positive selection and four ML approaches identified site 147 of TS as a likely target of positive selection. For DBAT, no sites were identified as unambiguous targets of positive selection.

In addition, we used a genetic algorithm approach (GA branch) that searches for an optimal model of lineage-specific evolution by assigning four unrestricted classes of d_N/d_S to lineages (Kosakovsky Pond and Frost 2005c). This approach allows an averaged model probability that d_N/d_S is greater than one along a specific lineage. Unlike branch site methods, GA branch does not need the user to select branches of interest to test, or test

Table 2 Likelihood values and parameter estimates for the *TS* (exons 1–4) gene

Model code	Log-likelihood	κ	α	β	Estimates of parameters	Positively selected sites[a]
M8[b] (Beta & ω)	−2263.58	2.29336	0.117898	2.70564	ω_s: 1.64644 prop(ω_s): 0.349296	111, 147, 193, 207, 214, 224, 261, 264, 274, 275, 276, 277, 295, 298, 328, 332, 343
M8a (null model)	−2265.83	2.22412	0.117898	2.70564	ω_s set to 1 prop (ω_s): 0.428767	Not allowed
M7 (Beta)	−2267.47	2.43666	0.309731	0.395431	–	Not allowed
MEC[c]	−2247.64	–	0.14519	1.20741	Rate (transition): 4.15165 Rate (transversion): 1.58359 f: 0.679	147, 193, 298, 332
SLAC	−2265.81	–	–	–	$\omega = 0.46851$ 95% CI: 0.37361–0.57848	193
FEL	–	–	–	–	0.21289 subs/nucleotide	147, 193
REL	–	–	–	–	0.14998 subs/nucleotide	147, 193

[a] Only sites with Ka/Ks > 1 where the 95% confidence interval is larger than 1 (i.e., the lower bound is larger than 1) are considered as significant

[b] M8: α and β are the shape parameters of the beta distribution. κ is the transition/transversion ratio. ω_s is the additional category representing positive selection. prop(ω_s) is the proportion of sites under selection

[c] MEC: f is the proportion of sites under no selection. Similar to PAML, the MEC model assumes a beta distribution with parameters α and β

Table 3 Likelihood ratio statistics ($\Delta\ell$) and AICc scores for tests of positive selection

Gene region	M8 vs. M8a (df = 1)		MEC vs. M8a		M8 vs. M7 (df = 2)		
	Log-likelihood	*P*	Log-likelihood	AICc	Log-likelihood	−2ΔlnL	*P*
TS, exons 1–4[a]	−2263.58/−2265.83	<0.05	−2247.64/−2265.83	4505.33/4539.69	−2263.58/−2267.47	7.78	<0.05
TS, exons 1–9[b]	−3456.62/−3463.89	<0.001	−3443.86/−3463.89	4336.73/4363.45	−3456.62/−3468.17	23.10	<0.001
DBAT[c]	−2514.47/−2516.19	>0.05	−2518.84/−2516.19	5047.72/5040.41	−2514.47/−2518.01	7.08	<0.05

$-2\Delta\ln L = 2(\ln L_{\text{alternative hypothesis}} - \ln L_{\text{null hypothesis}})$, x^2 distribution

[a] 19 sequences

[b] Nine sequences

[c] 18 sequences

$AICc = -2\log L + 2p\frac{N}{N-p-1}$, *L*, likelihood, *P*, no. of free parameters, *N*, the sequence length. The smaller the AICc value, the better the model explains the data

one branch at a time (which can lead to statisitcal instability or acceptance of poorly supported models), but rather mines the data for good-fitting models. In addition, inference based on multiple models (as opposed to a null-alternative pair) is more robust to model misspecification. Ninety-five percent confidence intervals (CIs) for the AICc (Kosakovsky Pond and Frost 2005c) for the best model (c-AIC for *TS* = 4,396.57, *DBAT* = 4,901.59) did not overlap with the AICc measure for the single-rate model (c-AIC for *TS* = 4,413.11, *DBAT* = 4917.01) for the two loci. For the *TS* locus, the lineages *T. cuspidata*, *T. mairei*, *T. contorta*, and *T. wallichiana* were placed into a d_N/d_S category of 1.681 (Fig. 4), with a model-averaged probability of 92.7%, 98.0%, 92.3%, and 93.3%, respectively; the lineages *T.* × *hunnewelliana*, *T. cuspidata* var. *nana*, and *T. wallichiana* var. *yunnanensis* were placed into a d_N/d_S category of 10,000 (infinity; all substitutions along a given short branch are non-synonymous), with a model-averaged probability of 93.1%, 92.2%, 91.6%, and 98.7%, respectively. However, the 95% CIs for individual branch estimates of d_N/d_S were significantly different from one in only two cases, i.e., *T. mairei* and *T. wallichiana* var. *yunnanensis*. We suggest that this variation is mediated by the diversity in the different loads and types of pathogens and herbivores that *Taxus* were exposed to as they radiated. On the contrary, for the *DBAT* locus, although three branches were placed into a d_N/d_S category of 10,000 (data not

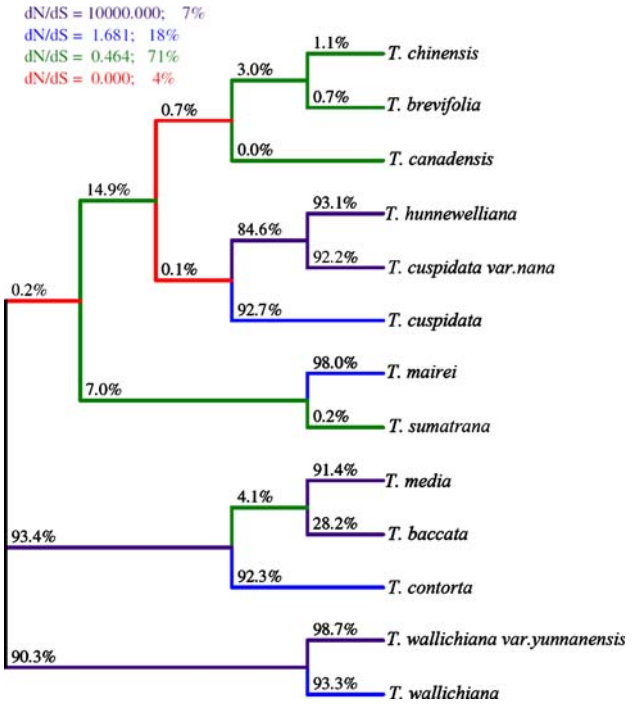

Fig. 4 Results from the genetic algorithm approach to detecting lineage-specific variation of TS in selection. The implemented HKY85 model was selected by HYPHY. Four unrestricted classes of d_N/d_S are assigned to branches (d_N/d_S values shown at the top left corner). Branch labels represent model averaged probabilities of $d_N/d_S > 1$ for the branch. Percentages for branch classes in the legend reflect the proportion of total tree length (measured in expected substitutions per site per unit time) evolving under the corresponding value of d_N/d_S. Branches in this tree are unscaled

shown), they failed to receive high model-averaged support for $d_N > d_S$.

Parallel amino acid substitutions, codon usage bias

The pattern of amino acid change based on ML ancestral sequence reconstruction provides further evidence that TS (and possibly DBAT) evolved under positive selection. Seven amino acid sites in TS and nine sites in DBAT changed independently to the same amino acid in 2 or more *Taxus* species (Table 4). For example, site 193 in TS changed from neutral and small residue threonine (T) to nonpolar isoleucine (I) in four different species. In DBAT, site 228 changed from neutral and small serine (S) to nonpolar and large phenylalanine (F) in two species. This change was classified as radical based on Grantham's distance, which takes into account amino acid size, hydrophobicity, charge, and polarity. Changes from a noncharged to a charged residue were also found in both enzymes. Changes of G55R and Y261S of TS, and G225V, S228F, A245E, and S294L of DBAT were not conservative or moderately conservative, as defined by changes in charge or by Grantham's distance. Such nonconservative changes have been found to occur much less frequently than expected under neutrality (Li et al. 1984). Thus, the nonconservative changes we observed seem more likely to have consequences for enzyme structure and/or function. Parallel evolution at the amino acid sequence level can be interpreted as evidence of adaptive evolution (Zhang 2003); consequently, sites that have changed in parallel are likely targets of selection in addition to those identified with the ML approach. These results also suggest that the number of available pathways of adaptive evolution may be constrained.

Several parameters related to codon usage bias were estimated to check whether synonymous mutations are selectively neutral. CBI is a measure for the deviation from the equal use of synonymous codons. CBI values range from 0 (uniform use of synonymous codons) to 1 (maximum codon bias). CBI values (*TS* 0.215–0.285, *DBAT* 0.236–0.258) were intermediate between uniform use of synonymous codons and maximum codon bias for both gene regions. Additionally, the ENCs were calculated, which may range from 20 (only one codon is used for each amino acid; i.e., the codon bias is maximum) to 61 (all synonymous codons for each amino acid are equally used; i.e., no codon bias). ENC values (*TS* 55.527–60.642, *DBAT* 55.146–56.68) were intermediate. C + G content at the second and third codon position was comparable to the overall C + G content of the total gene region (*TS* 0.47–0.479, *DBAT* 0.43–0.436), and ranged between 37.4 and 47.4%.

Discussion and conclusion

Species level polytomies may confound inferences of positive selection, due to the presence of recombination within loci (or between loci for concatenated datasets). Maximum likelihood methods implemented in PAML are often used to detect the action of positive selection on coding sequences. These methods are known to be sensitive to recombination; moderate to high levels of recombination can lead to an unacceptably high false positive rate (Anisimova et al. 2003). The increased false positive rate associated with recombination may result from the assumption that the rate of synonymous substitution is homogeneous across all sites (non-synonymous substitution rates are allowed to vary between codons), or from the use of an incorrect tree for some sites (Anisimova et al. 2003). Although lineage sorting in a deep ancestor has not been explicitly investigated as a source of error in PAML and related analyses, it may have confounding effects. We therefore combined several approaches, PAML models, MEC model, and codon-based maximum

Table 4 Parallel amino acid substitutions in TS and DBAT

Site	No. of independent changes	Starting amino acid class	Ending amino acid class	Charge changing	Grantham's distance	Type of change	Possible alternative amino acid substitutions
TS							
G55R	2	P	+	Y	125	MR	6
Y111F	3	P	NP	N	22	C	6
V147L	**2**	**NP**	**NP**	**N**	**32**	**C**	**5**
F168L	6	NP	NP	N	22	C	6
T193I	**4**	**P**	**NP**	**N**	**89**	**MC**	**5**
Y261S	7	P	P	N	144	MR	6
N298K	3	P	+	Y	94	MC	7
DBAT							
V57A	4	NP	NP	N	64	MC	5
H148R	7	+	+	N	29	C	7
S219T	2	P	P	N	58	MC	5
G225V	7	P	NP	N	109	MR	6
S228F	2	P	NP	N	155	R	6
A245E	7	NP	−	Y	107	MR	6
S294L	3	P	NP	N	145	MR	5
M352I	4	NP	NP	N	10	C	6
V422M	3	NP	NP	N	21	C	5

Amino acid types: +, positively charged; −, negatively charged; P, polar; NP, nonpolar. Grantham's distance between starting and ending amino acid (Grantham 1974). Types of change: C, conservative (Grantham's distance < 50); MC, moderately conservative (51–100); MR, moderately radical (101–150); R, radical (>150) (Li et al. 1984). Possible alternative amino acid substitutions: number of possible amino acid substitutions given the starting codon and a single nucleotide mutation. Rows in bold are sites identified as likely targets of positive selection

likelihood methods (SLAC, FEL, and REL) that can take recombination into account, to minimize inferential problems stemming from possible lineage sorting and recombination.

A comparison between *Taxus TS* and *DBAT* indicates that *TS* exons 1–4 have experienced greater selective pressure to change its amino acid composition than has *DBAT*. The models identified a number of positively selected sites in *TS* exons 1–4, and there was consensus between models for two of these sites. The present study thus provides the first insight into the positive selection exerted on the paclitaxel biosynthetic enzymes. This is also the first report of the positive selection of the secondary metabolism enzyme within a single genus, which complements examining evolutionary processes at multiple taxonomic levels. Terpene synthases are a mechanistically intriguing family of enzymes that catalyze complex, multistep reactions that are capable of generating hundreds of structurally diverse hydrocarbon and oxygenated scaffolds of biological and commercial importance. It was suggested that all plant terpene synthases, including TSs, share a common evolutionary origin (Trapp and Croteau 2001). The ancestral gene diverged in structure and function, by adaptive evolutionary processes, to yield the large superfamily of terpene synthases involved in secondary metabolic pathways. TS catalyzes the cyclization of geranylgeranyl diphosphate to taxa-4(5),11(12)-diene (Walker and Croteau 2001) and, in constructing the unique taxane skeleton, constitutes the committed step in the biosynthesis of paclitaxel and related taxoids. Although speculative, it is plausible that the positive selection and the parallel amino acid substitution, besides the gene organization, have been a driving force in the evolution of TS. For example, the changes of valine to leucine of site 147 and threonine to isoleucine of site 193 could make the N-terminal region more hydrophobic; the change of tyrosine to phenylalanine of site 111 could cause the loss of a tyrosine phosphorylation site (predicted by NetPhos 2.0). Such changes might affect the structural and/or catalytic function (although poorly studied) of CDIS domain encoded by *TS* exons 2–4 and subsequently affect the substrate binding motif and the active site domain at the C-terminal portion of the enzyme. Compared to the N-terminal region, the C-terminal active site including *TS* exons 10–13 remains highly conserved in organization and catalytic function (Trapp and Croteau 2001). TS (both native and recombinant) produces a small amount of taxadiene isomers ($\sim$6%; Williams et al. 2000) except the major product, taxa-4(5),11(12)-diene (94%). Whether the identified positive selection and the parallel amino acid substitution alter the product profile in the

respective *Taxus* species deserves further study. The findings of positive selection in the evolution of paclitaxel biosynthetic enzymes such as TSs are counterintuitive in the light of the supposed limited room for change in these molecules. The findings then support expectations of both high selection pressure acting on the various *Taxus* species within their unique habitats and significant changes in intensity and direction (kinds of pathogens and herbivores) resulting from changes in microhabitat and food.

Acknowledgments We thank the following experts for providing plant materials: YunFen Geng (YunNan Academy of Forestry, KunMing, China), YinKe Zhang (HangZhou Botanical Garden, China), Ron Determann (Atlanta Botanical Garden, GA, USA), Richard W. Spjut (World Botanical Associates, CA, USA), Robert G. Nicolson (Smith College, USA), Stephane Bailleul (Montreal Botanical Garden, Canada), Eric La Fountaine (University of British Columbia Botanical Garden, Canada), and James Stevenson (University of Oxford Botanic Garden, UK). We are grateful to Sergei L. Kosakovsky Pond and Leslie M. Turner (University of California, San Diego, USA) for suggestions in genetic algorithm and parallel amino acid substitutions, respectively, and to two anonymous reviewers for their critical comments. This study is supported by the National 973 Project (2007CB707802) of the Ministry of Science & Technology of China.

References

Anisimova M, Nielsen R, Yang Z (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164:1229–1236

Bishop JG (2005) Directed mutagenesis confirms the functional importance of positively selected sites in polygalacturonase inhibitor protein. Mol Biol Evol 22:1531–1534

Cheng Y, Nicolson RG, Tripp K, Chaw SM (2000) Phylogeny of taxaceae and cephalotaxaceae genera inferred from chloroplast matK gene and nuclear rDNA ITS region. Mol Phylogenet Evol 14:353–365

Doron-Faigenboim A, Pupko TA (2007) Combined empirical and mechanistic codon model. Mol Biol Evol 24:388–397

Grantham R (1974) Amino-acid difference formula to help explain protein evolution. Science 185:862–864

Hartmann T, Kutchan TM, Strack D (2005) Evolution of metabolic diversity. Phytochemistry 66:1198–1199

Koepp E, Hezari M, Zajicek J et al (1995) Cyclization of geranylgeranyl diphosphate to taxa-4(5),11(12)-diene is the committed step of Taxol biosynthesis in Pacific yew. J Biol Chem 270:8686–8690

Kosakovsky Pond SL, Frost SD (2005a) Not so different after All: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22:1208–1222

Kosakovsky Pond SL, Frost SD (2005b) Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics 21:2531–2533

Kosakovsky Pond SL, Frost SD (2005c) A genetic algorithm approach to detecting lineage-specific variation in selection pressure. Mol Biol Evol 22:478–485

Kress WJ, Wurdack KJ, Zimmer EA et al (2005) Use of DNA barcodes to identify flowering plants. Proc Natl Acad Sci USA 102:8369–8374

Li WH, Wu CI, Luo CC (1984) Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol 21:58–71

Li J, Davis CC, Tredici PD et al (2001) Phylogeny and biogeography of *Taxus* (Taxaceae) inferred from sequences of the internal transcribed spacer region of nuclear ribosomal DNA. Harvard Papers Bot 6:267–274

Liu Z, Bos JIB, Armstrong M et al (2005) Patterns of diversifying selection in the phytotoxin-like scr74 gene family of Phytophthora infestans. Mol Biol Evol 22:659–672

Morton BR (1993) Chloroplast DNA codon use: evidence for selection at the psbA locus based on tRNA availability. J Mol Evol 37:273–280

Nielsen R, Bustamante C, Clark AG et al (2005) A scan for positively selected genes in the genomes of Humans and Chimpanzees. PloS Biol 3:976–985

Posada D (2006) ModelTest Server: a web-based tool for the statistical selection of models of nucleotide substitution online. Nucleic Acids Res 34:W700–703

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574

Rozas J, Sánchez-DelBarrio JC, Messeguer X et al (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496–2497

Sharp PM (1991) Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position, and concerted evolution. J Mol Evol 33: 23–33

Spjut RW (2007a) Taxonomy and nomenclature of *Taxus*. J Bot Res Inst Texas 23:203–289

Spjut RW (2007b) A phytogeographical analysis of *Taxus* (Taxaceae) based on leaf anatomical characters. J Bot Res Inst Texas 23:291–332

Swanson WJ, Yang Z, Wolfner MF et al (2001) Positive darwinian selection in the evolution of mammalian female reproductive proteins. Proc Natl Acad Sci USA 98:2509–2514

Swofford DL (2002) PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Massachusetts

Trapp SC, Croteau RB (2001) Genomic organization of plant terpene synthases and molecular evolutionary implications. Genetics 158:811–832

van Rozendaal EL, Lelyveld GP, van Beek TA (2000) Screening of the needles of different yew species and cultivars for paclitaxel and related taxoids. Phytochemistry 53:383–389

Walker K, Croteau RB (2000) Molecular cloning of a 10-deacetylbaccatin III-10-O-acetyl transferase cDNA from Taxus and functional expression in *Escherichia coli*. Proc Natl Acad Sci USA 97:583–587

Walker K, Croteau RB (2001) Taxol biosynthetic genes. Phytochemistry 58:1–7

Wildung MR, Croteau RB (1996) A cDNA clone for taxadiene synthase, the diterpene cyclase that catalyzes the committed step of Taxol biosynthesis. J Biol Chem 271:9201–9204

Williams DC, Wildung MR, Jin AQ et al (2000) Heterologous expression and characterization of a "Pseudomature" form of taxadiene synthase involved in paclitaxel (Taxol) biosynthesis and evaluation of a potential intermediate and inhibitors of the multistep diterpene cyclization reaction. Arch Biochem Biophys 379:137–146

Wright F (1990) The "effective number of codons" used in a gene. Gene 87:23–29

Yang Z (2007) PAML 4: a program package for phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591

Yang Z, Kumar S, Nei M (1995) A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141:1641–1650

Yang Z, Nielsen R, Goldman N et al (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449

Zhang J (2003) Parallel functional changes in the digestive RNases of ruminants and colobines by divergent amino acid substitutions. Mol Biol Evol 20:1310–1317

Zwickl DJ (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. dissertation, The University of Texas at Austin