J Mol Evol (2006) 63:108–119
DOI: 10.1007/s00239-005-0249-2

JOURNAL OF MOLECULAR EVOLUTION

Molecular Evolution of a Small Gene Family of Wound Inducible Kunitz Trypsin Inhibitors in *Populus*

Nina M. Talyzina, Pär K. Ingvarsson

Umeå Plant Science Centre, Department of Ecology and Environmental Science, Umeå University, SE-90187 Umeå, Sweden

Received: 20 October 2005 / Accepted: 15 February 2006 [*Reviewing Editor:* Dr. Willie J. Swanson]

Abstract. Maximum likelihood models of codon substitutions were used to analyze the molecular evolution of a Kunitz trypsin inhibitor (*KTI*) gene family in *Populus* and *Salix*. The methods support previous assertions that the *KTI* genes comprise a rapidly evolving gene family. Models that allow for codon specific estimates of the ratio of nonsynonymous to synonymous substitutions (ω) among sites detect positive Darwinian selection at several sites in the *KTI* protein. In addition, branch-specific maximum likelihood models show that there is significant heterogeneity in ω among branches of the *KTI* phylogeny. In particular, ω is substantially higher following duplication than speciation. There is also evidence for significant rate heterogeneity following gene duplication, suggesting different evolutionary rates in newly arisen gene duplicates. The results indicate uneven evolutionary rates both between sites in the *KTI* protein and among different lineages in the *KTI* phylogeny, which is incompatible with a neutral model of sequence evolution.

Key words: Herbivore defense — Kunitz trypsin inhibitor — Maximum likelihood — *Populus* — Positive selection — Gene duplication

Introduction

Interactions with herbivores and pathogens are known to have significant effects on plant fitness. Consequently, plants have developed various defense mechanisms in order to prevent, decrease, and repel herbivore and pathogen attack (Karban and Baldwin 1997; Agrawal et al. 1999). Plant genomes often contain a large number of different defense genes. Since gene duplication is widely regarded as a mechanism that can lead to acquisition of new gene functions, it has been suggested that gene duplication is an important source for novel genes involved in plant defense (Michelmore and Meyers 1998; Mondragon-Palomino et al. 2002; Moore and Purugganan 2005).

Interest in the evolutionary fate of gene duplications traces back to the pioneering work of Ohno (1970). Ohno assumed that one copy is redundant following the duplication event and would eventually turn into a pseudogene due to the accumulation of deleterious mutations (nonfunctionalization) or could fortuitously turn into a novel gene (neofunctionalization). Since deleterious mutations are more likely than beneficial ones, the prevailing thought has been that most duplicate gene copies are destined to become pseudogenes. Over the last decade it has become apparent, however, that positive selection can play an important role in preserving some gene copies (Walsh 1995; Lynch et al. 2001; Moore and Purugganan 2003). For instance, Moore and Purugganan (2003) found that in three recent duplicated gene pairs in *A. thaliana*, patterns of nucleotide polymorphism were consistent with positive selection acting on either one or both gene copies and thus support neofunctionalization at these loci. Such retention and diversification of duplicate copies through positive selection might be particular likely in plant defense genes (Michelmore and Meyers 1998; Moore and Purugganan 2005). Segmental duplication followed by positive selection has resulted in a sub-

Correspondence to: Pär K. Ingvarsson; *email:* pelle@wallace.emg.umu.se

stantial increase in the number of pathogen recognition loci (*R* loci) in *Arabidopsis thaliana*, even though this gene family has rates of duplication and mutation similar to those of other gene families (Leister 2004; Meyers et al. 2005). Similarly, in a tandemly arrayed trypsin inhibitor family in *Arabidopsis thaliana*, significant variation between loci in patterns of gene expression is likely a response to divergent natural selection from a diverse set of herbivores (Clauss and Mitchell-Olds 2004).

Plants actively defend themselves against pathogen and herbivores using several inducible defense mechanisms (Karban and Baldwin 1997). One such class of inducible herbivore defense genes is protease inhibitors (PIs). The function of PIs in herbivore defense is well established (Ryan 1990; Heitz et al. 1999; Lee et al. 1999). Plant PIs are proteins that function as specific substrates for proteolytic enzymes in the digestive tracts of herbivores (Ryan 1990; Constabel 1999; Haq et al. 2004). However, a PI is not cleaved by a protease like normal substrates are but, rather, form a stable complex that limits or completely inhibits the proteolytic activity of the protease (Ryan 1990; Constabel 1999; Haq et al. 2004). This results in a much reduced proteolysis in the digestive tract of herbivores, a lack of available amino acids, and ultimately lowered growth rates or starvation of the herbivore (Ryan 1990; Constable 1999).

In the genus *Populus*, a number of Kunitz trypsin inhibitors (*KTI*) are among the most up-regulated genes shortly after damage by herbivores (Bradshaw et al. 1990; Saarikoski et al. 1996; Haruta et al. 2001; Christopher et al. 2004). These genes have been shown to belong to a small gene family with amino acid sequence similarities ranging from 28% to over 90% (Haruta et al. 2001; Christopher et al. 2004). While the transcription of all *KTI* genes is highly upregulated following herbivory in *Populus*, the spatial and temporal patterns of induction of the five genes are quite different and distinct, both locally, at the wound site, and systemically throughout the plant, suggesting that some functional diversification has taken place among the gene family members (Haruta et al. 2001; Christopher et al. 2004).

In this paper, we apply a phylogenetic perspective to the evolution of the *KTI* gene family in *Populus* in an attempt to analyze historical selection pressures on individual genes in the gene family. We use maximum-likelihood methods (Yang 1998) to test for the action of positive selection throughout the diversification of the *KTI* gene family. Special attention is paid to patterns of molecular evolution operating shortly after gene duplication and following speciation. We are particularly interested assessing relative roles of positive natural selection and neutral genetic drift in driving early evolution of newly arisen gene duplicates in the Kunitz PI gene family in *Populus*.

Materials and Methods

Kunitz trypsin inhibitor sequences were gathered from 10 species of *Populus* and from 3 species of *Salix* that were used as outgroup sequences. Five members of the Kunitz *TI* gene family in Salicaceae have been recognized previously: *KTI1*, *KTI2*, *KTI3* (Haruta et al. 2001) *KTI*4, and *KTI*5 (Christopher et al. 2004). *KTI4* and *KTI5* represent diverged members of the Kunitz class of PIs (25%–35% amino acid similarity with *KTI1-KTI3* [Christopher et al. 2004]), but the remaining gene copies appear to form a small gene family with between 60% and 90% amino acid similarity (Haruta et al. 2001). In the present analyses we focused on the family members *KTI1*, *KTI2* and *KTI3*, and two copies, *KTI6* and *KTI7*, that were identified during the present investigation. *KTI6* is identical to the *TIH-2.1* gene earlier described in *P. nigra* (GenBank reference AB081325), whereas *KTI7* appears to be previously undescribed *KTI* copy. The presence of these two genes in the *Populus* genome was verified through BLAST searches of the assembled *P. trichocarpa* genome sequence.

The complete phylogenetic tree was constructed using a total sample of 24 sequences, representing 5 paralogous *KTI* gene copies of *Populus* and 3 sequences of *Salix*. Six *KTI* sequences (from *P. alba* [*KTI2*], *P. nigra* [*KTI7*], *P. tremula* [*TI6*], *P. alba* [*TI3*], *S. myrsinifolia*, and *S. triandra*) were obtained by PCR amplification from silica-dried leaf tissue using primers targeting conserved sites in the various *KTI* sequences. These six new sequences have been deposited in the EMBL database under accession numbers AM072338–AM072343. Remaining sequences included in the study were either obtained from existing GenBank records (Table 1) or by blast searches of the assembled *Populus trichocarpa* genome sequence (http://genome.jgi-psf.org/Poptr1/Poptr1.home.html).

To amplify *KTI* sequences total DNA was extracted using a Qiagen DNeasy Plant Mini Kit. PCR conditions were 94°C (30 s)/48°C (30 s)/72°C (60 s), followed by 72°C (2 min). PCR products were cleaned using the Viogene PCR-M Clean up system (spin method) and cloned into the pCR2.1 vector using a TA-Cloning Kit from Invitrogen (Carlsbad, CA). Three to nine clones of each PCR fragment were sequenced using BigDye chemistry (Applied Biosystems Inc., Foster City, CA) on an ABI377 automated sequencer at the Umeå Plant Science Centre sequencing facility. Sequences were manually verified and contigs assembled using Sequencer 4.1 (Gene Codes Corp.).

Multiple alignments were made using Clustal W using default parameters, followed by adjustments by hand in BioEdit (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). A neighbor-joining (NJ) phylogenetic tree was constructed in MEGA 2.1 (http://www. megasoftware.net) using Kimura two-parameter distances and pairwise deletion of gaps. We also generated a phylogenetic tree by the maximum likelihood (ML) method implemented in PAML. The ML tree was generated using a random addition of sequences and Kimura two-parameter distances with the transition-transversion ratio estimated from the data and no rate variation among sites. This tree had an identical topology to the NJ tree and was therefore used for all analyses described below. The robustness of the tree was tested by boostraping the NJ tree in MEGA 2.1 and the ML tree in PAUP. The results were virtually identical and only the results from the NJ analysis are presented.

Estimates of nonsynonymous (dN) and synonymous substitutions (dS) and their ratio ω = dN/dS, were obtained using the codeml program of the PAML (v3.14) package (Yang 1998; http://abacus.gene.ucl.ac.uk/software/paml.html). We applied three different types of analyses to our data: one that allows for site-to-site variation in ω (Yang and Nielsen 1998; Yang et al. 2000), another that allows ω to vary among branches (Yang 1998), and one that allows both site and branch ω variation (Yang and Nielsen 2002).

Table 1. Sequences Analyzed

Name	GenBank accession no.
Populus alba TI2	AM072338
P. tremuloides TI2	AF349442
P. tremula TI2	AJ936976
P. tremuloides TI1	AF349441
P. tremula TI1	AF937139
P. nigra TIH-1.2	AB081325
P. xgenerosa win 3.12	L11233
P. trichocarpa gwin 3	X15516
P. nigra TI7	AM072340
P. trichocarpa TI7	From *P. trichocarpa* genome sequence
P. trichocarpa TI6	From *P. trichocarpa* genome sequence
P. nigra TIH-2.1	AB081325
P. tremula TI6	AM072343
P. tremula TI3	AJ843713
P. tomentiqlandulosa TI3	CF936929
P. tremuloides TI3	AF349443
P. alba TI3	AM072339
P. euphratica TI3	AF779785
P. trichocarpa TI3	From *P. trichocarpa* genome sequence
P. balsamifera TI3	AY378088
P. euramericana TI3	CX178935
Salix myrsinifolia	AM072341
S. triandra	AM072342
S. viminalis Swin-1.1	U29090

The PAML package implements several different models dealing with site-to-site variation in ω and each of these models has advantages and disadvantages. Yang et al. (2000) recommended that several different models should be applied when analyzing real data. Therefore, to test for variation among sites in ω we used the one-ratio model M0, a suite of site-specific models: M1 (neutral), M2 (selection), M3 (discrete), M7 (ß), M8 (ß and ω) (Yang 1998; Yang and Nielsen 2002; Yang et al. 2000), and a modification of the M8 model (M8A, Swanson et al. 2003). Branch-specific estimates of ω were obtained using either one ratio per branch or two independent ratios per branch (Yang 1998; Yang and Nielsen 2002). Finally, to specifically study which sites were under selection following gene duplication, we used models that allow for both branch and site variation (models A and B of Yang and Nielsen [2002]), treating lineages resulting from gene duplication events as "foreground lineages" and speciation branches as "background lineages" (*sensu* Yang and Nielsen 2002). In all analyses, nested models were compared using likelihood ratio tests (LRTs). Codon frequencies were estimated using average nucleotide frequencies at the three codon positions (option CodonFreq = 2 in codeml). In addition, rate variation among sites was ignored and the transition-transversion ratio (κ) was estimated from the data.

To test for rate variation between duplicate copies following gene duplication we performed relative rate tests using the method of Tajima (1993), as implemented in MEGA 2.1. We tested rate differences for three independent duplications; the *KTI1/KTI2* duplication in *P. tremula* using *P. trichocarpa gwin3* as an outgroup, the *KTI6/KTI7* duplication in *P. trichocarpa* using *P. trichocarpa TI3* as an outgroup sequence, and finally, the lineages ancestral to *KTI1/KTI2* and *KTI6/KTI7* duplications with the ancestral *KTI3* sequence as an outgroup. Ancestral sequences for the last analysis were reconstructed using the baseml program in PAML.

Results

In the phylogenetic tree constructed from the complete data set (Fig. 1) sequences from *Salix* cluster together on a separate branch. They also cluster with high bootstrap support with *KTI3* from *Populus*, suggesting that the *Salix* sequences represent orthologous copies of the *KTI*3 gene in *Populus*. In addition, the four remaining *KTI*s in *Populus* form separate clades well supported by high bootstrap values, confirming our assumption that they represent paralogous *KTI* copies.

To investigate whether natural selection has been a significant force during the diversification of the *KTI* gene family, we first focused on models that allow for variation among sites in rates of synonymous and nonsynonymous substitutions but do not take variation among different branches of the phylogenetic tree into account. Under the simplest model, which allows for only a single ω ratio across all sites (M0), the ML estimate of ω = 1.017 (Table 2). This estimate is statistically indistinguishable from ω = 1, the expected value under a completely neutral model of sequence evolution. However, a model allowing for variation among sites (M3; Tables 2 and 3) provides a significantly better fit to the data, showing that there is variation among sites in strength of selection (Tables 2 and 3). The M3 model identifies two classes of sites with ω > 1: roughly 51% of the sites have an average ω = 1.61, whereas an additional 7% have ω = 5.23. However, as noted by Anisimova et al (2001), M0 is a highly unrealistic model and the M0–M3 comparison thus really provides a test for variation in ω among sites rather than variation in strength of selection among sites. A more stringent test for the presence of positive selection is a comparison of models M7 and M8 (Anisimova et al. 2001). The LRT comparing M7 and M8 provides strong evidence for positive selection (Table 3). Similar results (Table 2) were obtained using models M1 and M2 and models M8A and M8 (Swanson et al. 2003). Under model M8 roughly 25% of the sites have ω = 2.92 and 14 sites are identified as targets of selection with a posterior probability of ω > 1 greater than 0.95 using the Bayes empirical Bayes method (Yang et al. 2005). Model M8 provides higher accuracy, but also lower power, than model M3 analysis. Model M2 has similar accuracy as model M8 and estimates that roughly 23% of the sites have been under positive selection, with an average ω = 2.97. Model M2 estimates that roughly 23% of sites have been under positive selection, with an average ω = 2.97, and identifies 13 sites as likely targets of positive selection with a posterior probability >0.95 (Table 2). There is a large degree of overlap in the sites identified from models M2 and M8; all 13 sites identified by model M2 are also identified by model M8.

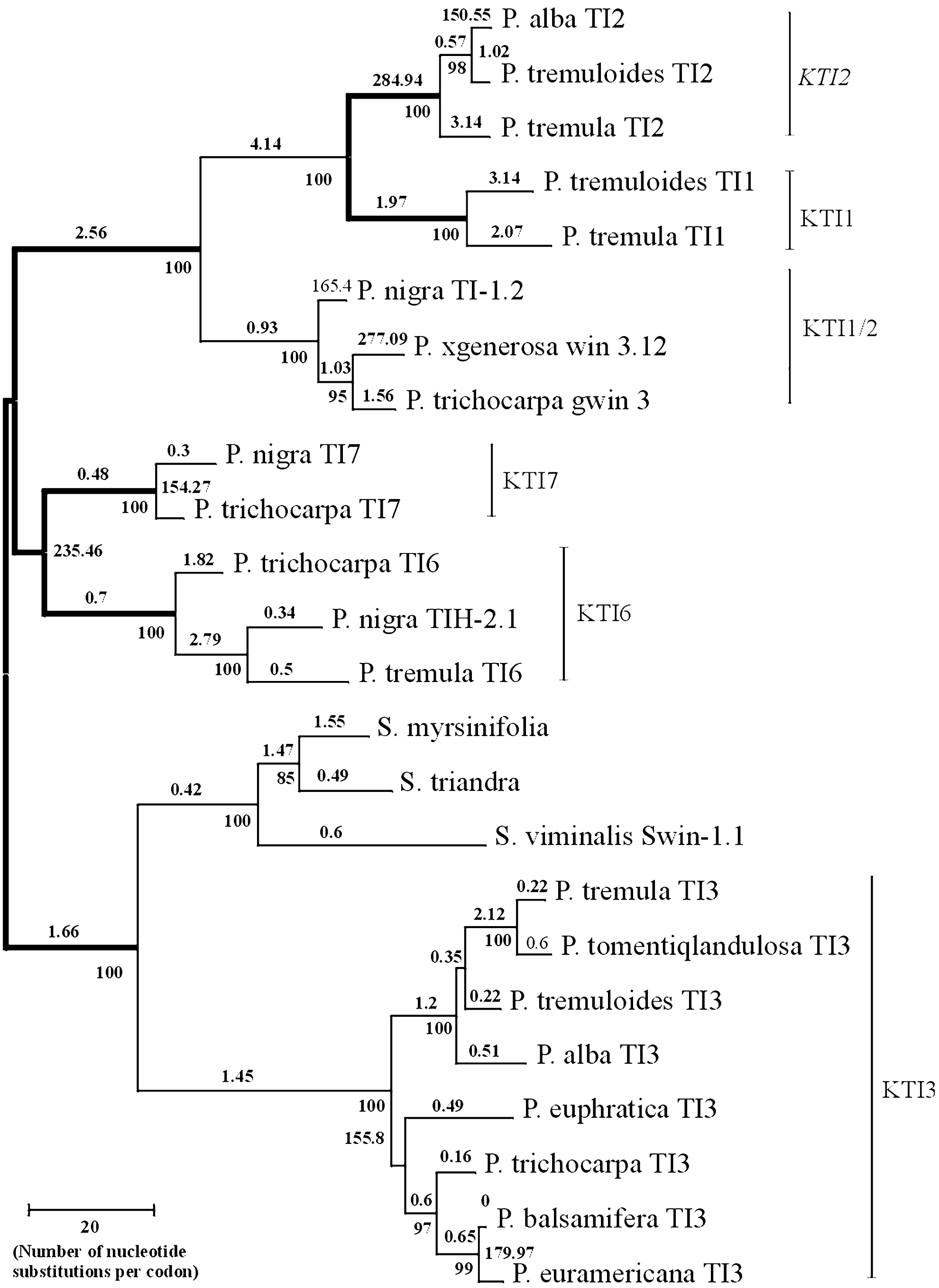

Fig. 1. Neighbor-joining phylogeny of *KTI* genes from Salicaceae; branch length corresponds to number of nucleotide differences. The tree topology is used to fit all maximum likelihood models considered in the paper. Numbers above branches are ω ratios estimated using the M1 NS0 model (Table 1) and numbers below branches represents bootstrap support from 2000 replicates (only bootstrap values >80% are shown for clarity). Duplication branches are denoted by bold lines (see text for further details).

The alignment in Fig. 2 has several sites with gaps. PAML treats gaps as ambiguities and this may be inappropriate for analyzing alignments with long indels or where less than half of the sequences in an alignment have residues in these long gaps. We therefore repeated all analyses with gapped sites excluded (data not shown). Results from these analyses are consistent with the results obtained using the complete data set. Ignoring gapped sites, model M8 shows that 27.2% of the sites have ω > 1 and the Bayes empirical Bayes method identifies 10 sites with >95% posterior probability. These 10 sites are a subset of the sites identified using the full data set (Table 2), where 14 sites were identified. The remaining four sites are absent from the gap-free data set because one or more sequences have gaps at those sites. Of these four sites, only one (site 205) has more than two sequences with gaps. With the possible

Table 2. Estimates for the *Populus*/*Salix* Data

Model	n_p	L	Estimates of parameters	Positively selected sites
M0: one ratio	47	−4380.14	$\omega = \mathbf{1.017}$	None
Site-specific models				
M1: neutral	48	−4304.22	$p_0 = 0.375$, $p_1 = 0.625$	Not allowed
M2: selection	50	−4273.61	$p_0 = 0.329$, $p_1 = 0.437$, $p_2 = 0.234$, $\omega_2 = \mathbf{2.965}$	51, 72, 117, 123, 195 ($p > 0.99$); 43, 57, 87, 118, 137, 204, 205, 216 ($p > 0.95$)
M3: discrete				
($K = 2$)	48	−4283.16	$p_0 = 0.487$, $p_1 = 0.513$, $\omega_0 = 0.209$, $\omega_1 = \mathbf{2.172}$	17, 24, 37, 40, 43, 46, 49, 51, 55, 57, 58, 64, 65, 68, 71, 72, 73, 81, 85, 87, 88, 98, 109, 113, 117, 118, 123, 125, 135, 136, 137, 139, 151, 154, 156, 160, 162, 164, 174, 194, 195, 197, 204, 205, 206, 208, 216, 217, 218, 219, 220, 223 ($p > 0.99$)
($K = 3$)	51	−4270.25	$p_0 = 0.413$, $p_1 = 0.515$, $p_2 = 0.072$, $\omega_0 = 0.133$, $\omega_1 = \mathbf{1.609}$, $\omega_2 = \mathbf{5.23}$	17, 24, 31, 32, 37, 40, 43, 46, 47, 49, 51, 52, 53, 55, 56, 57, 58, 64, 65, 68, 71, 72, 73, 74, 81, 84, 85, 87, 88, 98, 102, 108, 109, 113, 117, 118, 123, 125, 134, 135, 136, 137, 139, 151, 154, 156, 160, 162, 163, 164, 174, 180, 194, 195, 197, 198, 202, 203, 204, 205, 206, 208, 216, 217, 218, 219, 220, 223 ($p > 0.99$)
M7: β, neutral	48	−4308.03	$a = 0.009$, $b = 0.005$	Not allowed
M8: β & ω, selection	50	−4274.32	$p_0 = 0.748$, ($p_1 = 0.252$), $a = 0.16$, $b = 0.115$, $\omega = \mathbf{2.921}$	51, 72, 117, 123, 195, 216 ($p > 0.99$) 43, 57, 58, 87, 118, 137, 204, 205 ($p > 0.95$)
M8A: ω = 1, neutral	49	−4304.26	$p_0 = 0.376$, ($p_1 = 0.624$), $a = 7.389$, $b = 99$	Not allowed
Branch-specific models				
One ratio per branch	91	−4344.05	See Fig. 1	Not allowed
Two independent ratios	48	−4375.52	$\omega_{\mathrm{speciation}} = 0.863$, $\omega_{\mathrm{duplication}} = \mathbf{1.861}$	Not allowed
Branch-site models				
Model A	50	−4286.23	$p_0 = 0.346$, $p_1 = 0.516$, $p_2 + p_3 = 0.138$, $\omega_2 = \mathbf{7.571}$	Sites for foreground: 72, 73, 136 ($p > 0.99$)
Model B	51	−4268.51	$p_0 = 0.473$, $p_1 = 0.437$, $p_2 + p_3 = 0.09$, $\omega_0 = 0.189$, $\omega_1 = \mathbf{1.99}$, $\omega_2 = \mathbf{10.677}$	Sites for all lineages: 64, 156 ($p \geq 0.99$); 31, 32, 37, 46, 47, 49, 68, 81, 87, 102, 108, 109, 113, 151, 154, 160, 162, 163, 164, 180, 194, 197, 202, 204, 217, 220 ($p > 0.95$) Sites for foreground: 72, 73 ($p > 0.99$); 136 ($p > 0.95$)

Note. L, log likelihood; n_p, number of parameters; ω, ratio of synonymous/non-synonymous substitutions; *p*, proportion of sites estimated to be under positive selection with $\omega > 1$, *K*, number of classes of sites; a, b, parameters of the beta distribution. Parameters indicating positive selection are in bold.

Table 3. Likelihood ratio test statistics for testing hypotheses

Hypothesis tested	Models compared	Likelihood ratio tests (LRT)
Variation among sites, positive selection	M0–M3 ($K = 3$)	$2\Delta_L = 219.78$, *df* = 4, $p < 0.001$
	M7–M8	$2\Delta_L = 67.42$, *df* = 2, $p < 0.001$
	M1–M2	$2\Delta_L = 61.22$, *df* = 2, $p < 0.001$
	M8A–M8	$2\Delta_L = 29.94$, *df* = 1, $p < 0.001$
Variation of ω among branches	M0: one ratio per branch	$2\Delta_L = 72.18$, *df* = 44, $0.001 < p < 0.005$
$\omega_{\text{duplication}} \neq \omega_{\text{speciation}}$	M0 (single ω ratio): two independent ratios	$2\Delta_L = 9.24$, *df* = 1, $0.001 < p < 0.005$
Free ω in all branches vs. speciation/duplication	Two independent ratios: one ratio per branch	$2\Delta_L = 62.94$, *df* = 43, $0.025 < p < 0.05$
Positive selection at specific sites following duplication	M1–Model A	$2\Delta_L = 35.98$, *df* = 2, $p < 0.001$
	M3 ($K = 2$)–Model B	$2\Delta_L = 29.3$, *df* = 3, $p < 0.001$

Note. See Table 2 for a description of the different models. $2\Delta_L$, log likelihood difference between models.

exception of site 205, missing data thus do not appear to inflate the number of sites under positive selection.

To determine where in the *KTI* protein the likely targets of selection are located, we mapped structural features onto the protein sequence alignment of the *KTI* genes using the crystal structure of the Kunitz-type trypsin inhibitor from soybean (*STI*) (Fig. 2) (Song and Suh 1998). *KTI* proteins contain a predicted signal peptide consisting of 25–28 amino acids (Nielsen et al. 1997), which is cleaved off to form the active PI (Song and Suh 1998; see also Fig. 2). Mature *KTI* proteins consist of 10–12 anti-parallel β-strands forming the β-trefoil fold that is characteristic of Kunitz-type PIs (Song and Suh 1998). The proteins also contain four conserved Cys residues that form two disulfide bridges that act to stabilize the mature protein (Fig. 2). The mature protein carries a reactive region situated on an external loop, located between amino acid 92 and amino acid 100 in the alignment of the complete data set (Fig. 2). The reactive loop plays a vital role during the formation of an inhibitor-enzyme complex that ultimately leads to the inactivation of the protease. Amino acid composition of the reactive loop is known to have a large influence on the inhibitory function of PIs (Song and Suh 1998; Ravichandran et al. 1999). We mapped sites that were identified as being under the influence of positive selection onto the sequence alignment (Table 2, Fig. 2). These sites were primarily located in loop regions connecting the different β–strands (Fig. 2). To our surprise, none of the sites identified as a likely target of strong positive selection, using the stringent tests of models M2 and M8 (Table 2), were located in the reactive loop.

To test whether patterns of natural selection differ following gene duplication or speciation, we implemented a model in PAML that allows for two ω values, one for branches that result from speciation events and another for branches that result from duplication events. The results show that duplication branches are characterized by a higher ω value ($\omega = 1.86$) than speciation branches ($\omega = 0.86$; Table 2). A LRT comparing the duplication/speciation model to the model allowing for a single ω-value shows that the duplication/speciation model provides a significantly better fit to the data (Table 3). Finally, a model allowing for completely free variation of ω ratios among all lineages in the *KTI* phylogeny provides a marginally better fit to the data, compared to the duplication/speciation model (Table 3, Fig. 1), suggesting that there is additional among-branch variation that is not accounted for by the simple speciation/duplication contrast. Such additional rate variation could be the result of episodic selection during the diversification of the *KTI* gene family.

Finally, we applied models A and B of Yang and Nielsen (2002), which allow ω for variation both among sites and between selected "foreground" and "background" branches. We marked all newly duplicated lineages as foreground and all speciation branches as background. In model A, $\omega = 7.51$ for foreground branches, suggesting strong positive selection in these lineages. Model A also identifies three sites (72, 73, and 136) as targets of positive selection in lineages following duplication events (Table 2). Under model B, sites under positive selection were found across all lineages ($\omega = 1.99$) and therefore also in foreground lineages ($\omega = 10.68$). For the LRT test model A is compared to the site-specific model M1, in which ω is fixed at 1, whereas model B is compared to M3 (discrete; $K = 2$). The LRT test shows that both model A and model B provide a better fit to the data than the site-specific models (Table 3). These results support our previous conclusions that positive selection has been pervasive following gene duplication.

The *KTI1* and *KTI2* paralogues are more than 90% identical at the nucleotide level and thus appear to be the result of a relatively recent gene duplication event. Intraspecific polymorphism data from *P. tremula* show several signs suggesting that both *KTI1* and *KTI2* have been under the influence of positive selection (Ing-

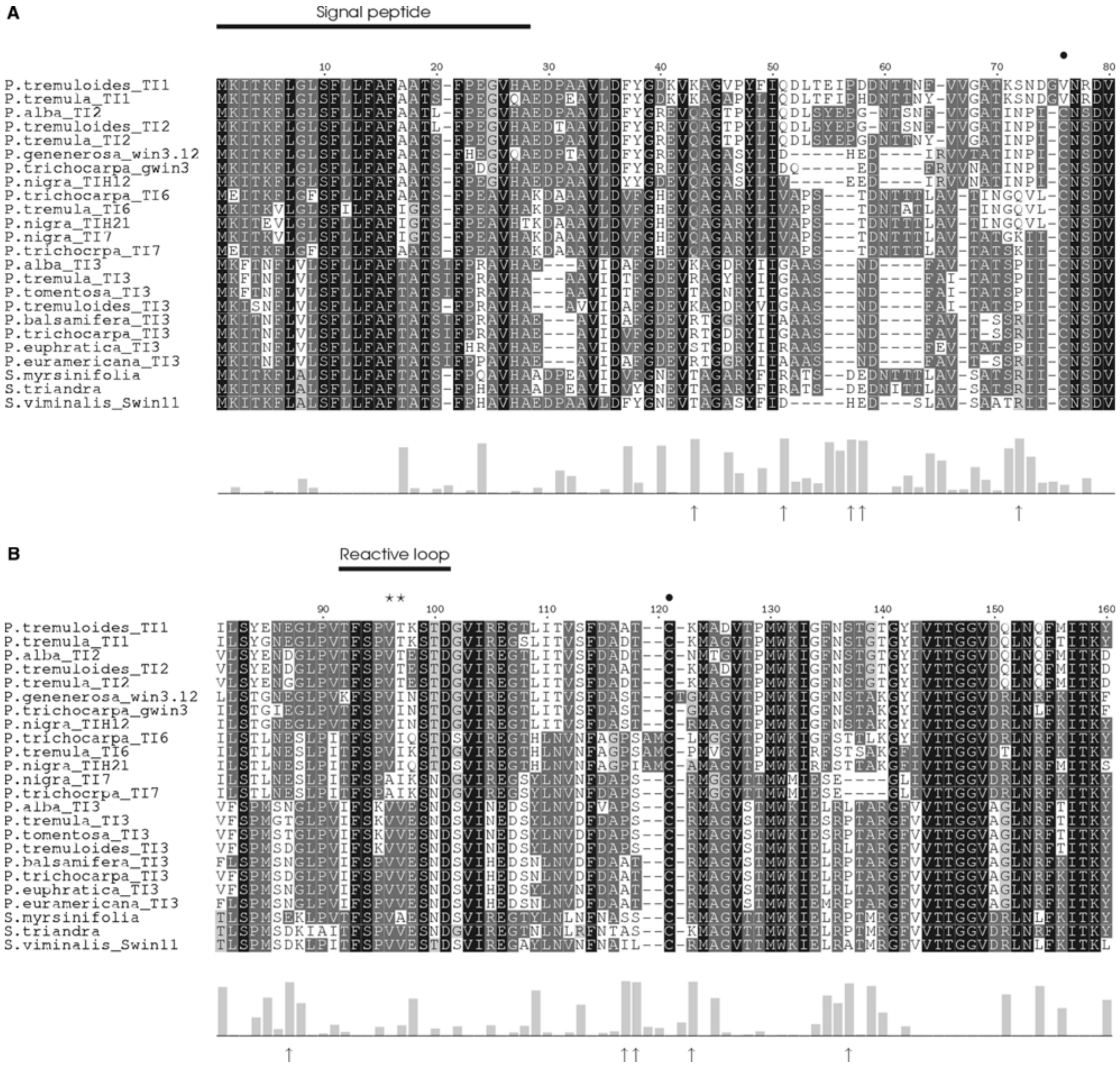

Fig. 2. Protein alignment of Kunitz trypsin inhibitor sequences from Salicaceae, showing conserved (black) and similar (gray) amino acid residues. Sequence gaps are shown by dashes. Structural features of the protein are shown above the alignment; stars indicate reactive-site residues P1 and P1′, which are crucial for protein inhibitory function (Song and Suh 1998); filled circles indicate conserved Cys residues forming disulfide bonds (Song and Suh 1998). Bars below alignment show a Bayesian posterior probability of ω > 1 for each site (see text for further details). Arrows indicate sites with $p_{\omega>1} > 0.95$. The figure was created using the TeXshade package (Beitz 2000).

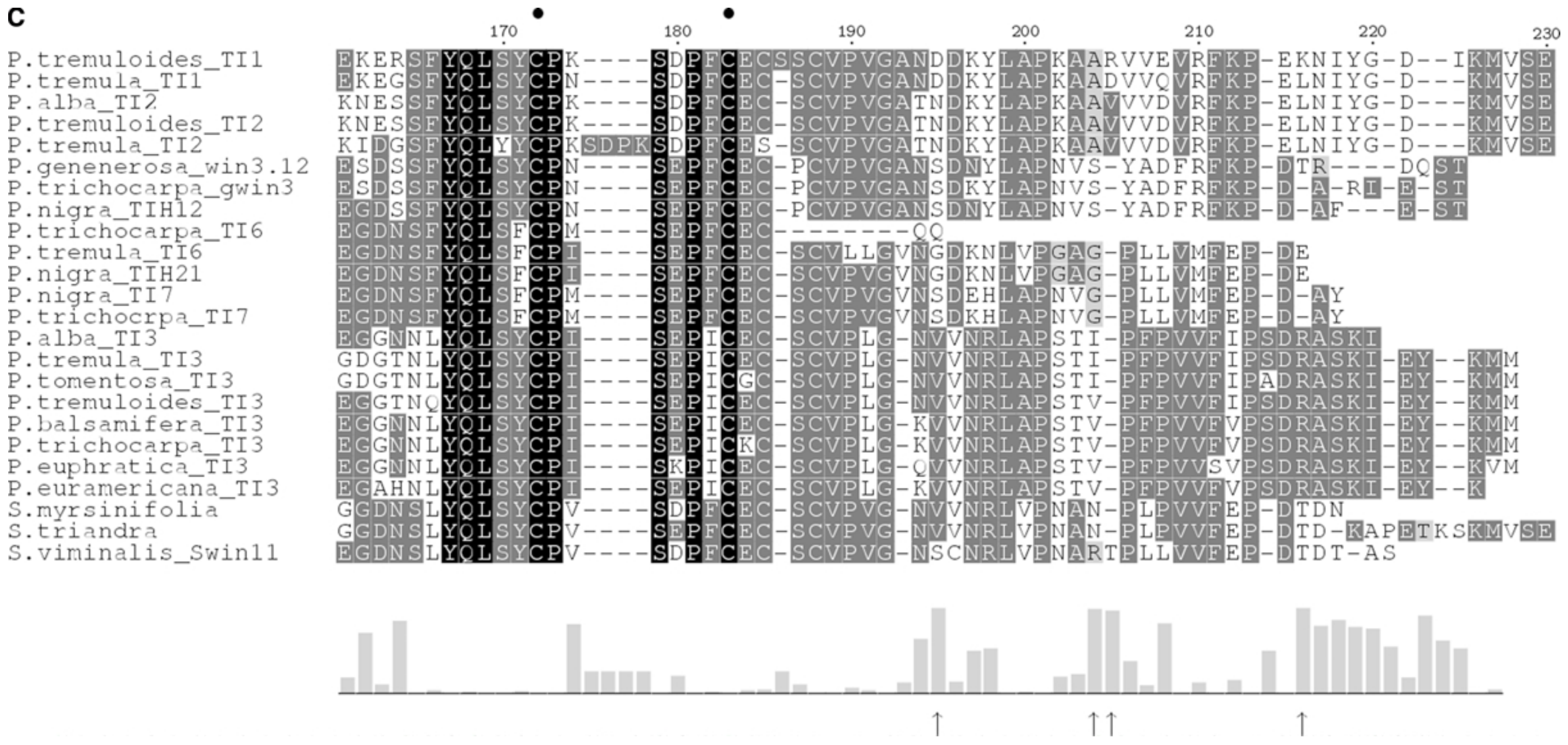
C
170
180
190
200
210
220
230
P.tremuloides_TI1 EKERSFYQLSYCPK----SDPFCECSSCVPVGANDDKYLAPKAARVVEVRFKP-EKNIYG-D--IKMVSE
P.tremula_TI1 EKEGSFYQLSYCPN----SDPFCEC-SCVPVGANDDKYLAPKAADVVQVRFKP-ELNIYG-D---KMVSE
P.alba_TI2 KNESSFYQLSYCPK----SDPFCEC-SCVPVGATNDKYLAPKAAVVVDVRFKP-ELNIYG-D---KMVSE
P.tremuloides_TI2 KNESSFYQLSYCPK----SDPFCEC-SCVPVGATNDKYLAPKAAVVVDVRFKP-ELNIYG-D---KMVSE
P.tremula_TI2 KIDGSFYQLYYCPKSDPKSDPFCES-SCVPVGATNDKYLAPKAAVVVDVRFKP-ELNIYG-D---KMVSE
P.genenerosa_win3.12 ESDSSFYQLSYCPN----SEPFCEC-PCVPVGANSDNYLAPNVS-YADFRFKP-DTR----DQST
P.trichocarpa_gwin3 ESDSSFYQLSYCPN----SEPFCEC-PCVPVGANSDKYLAPNVS-YADFRFKP-D-A-RI-E-ST
P.nigra_TIH12 EGDSSFYQLSYCPN----SEPFCEC-PCVPVGANSDNYLAPNVS-YADFRFKP-D-AF---E-ST
P.trichocarpa_TI6 EGDNSFYQLSFCPM----SEPFCEC--------QQ
P.tremula_TI6 EGDNSFYQLSFCPI----SEPFCEC-SCVLLGVNGDKNLVPGAG-PLLVMFEP-DE
P.nigra_TIH21 EGDNSFYQLSFCPI----SEPFCEC-SCVPVGVNGDKNLVPGAG-PLLVMFEP-DE
P.nigra_TI7 EGDNSFYQLSFCPM----SEPFCEC-SCVPVGVNSDEHLAPNVG-PLLVMFEP-D-AY
P.trichocrpa_TI7 EGDNSFYQLSFCPM----SEPFCEC-SCVPVGVNSDKHLAPNVG-PLLVMFEP-D-AY
P.alba_TI3 EGGNNLYQLSYCPI----SEPICEC-SCVPLG-NVVNRLAPSTI-PFPVVFIPSDRASKI
P.tremula_TI3 GDGTNLYQLSYCPI----SEPICEC-SCVPLG-NVVNRLAPSTI-PFPVVFIPSDRASKI-EY--KMM
P.tomentosa_TI3 GDGTNLYQLSYCPI----SEPICGC-SCVPLG-NVVNRLAPSTI-PFPVVFIPADRASKI-EY--KMM
P.tremuloides_TI3 EGGTNQYQLSYCPI----SEPICEC-SCVPLG-NVVNRLAPSTV-PFPVVFIPSDRASKI-EY--KMM
P.balsamifera_TI3 EGGNNLYQLSYCPI----SEPICEC-SCVPLG-KVVNRLAPSTV-PFPVVFVPSDRASKI-EY--KMM
P.trichocarpa_TI3 EGGNNLYQLSYCPI----SEPICKC-SCVPLG-KVVNRLAPSTV-PFPVVFVPSDRASKI-EY--KMM
P.euphratica_TI3 EGGNNLYQLSYCPI----SKPICEC-SCVPLG-QVVNRLAPSTV-PFPVVSVPSDRASKI-EY--KVM
P.euramericana_TI3 EGAHNLYQLSYCPI----SEPICEC-SCVPLG-KVVNRLAPSTV-PFPVVFVPSDRASKI-EY--K
S.myrsinifolia GGDNSLYQLSYCPV----SDPFCEC-SCVPVG-NVVNRLVPNAN-PLPVVFEP-DTDN
S.triandra GGDNSLYQLSYCPV----SEPFCEC-SCVPVG-NVVNRLVPNAN-PLPVVFEP-DTD-KAPETKSKMVSE
S.viminalis_Swin11 EGDNSLYQLSYCPV----SDPFCEC-SCVPVG-NSCNRLVPNARTPLLVVFEP-DTDT-AS

Fig. 2. Continued

Table 4. Relative rate tests of duplicated *KTI* genes

Sequence 1	Sequence 2	Outgroup	Lineage-specific substitutions (Seq1/Seq2/Out)	Relative rate test[a]
P. tremula TI1	*P. tremula TI2*	*P. trichocarpa gwin3*	28/11/60	$\chi^2 = 7.05$, $p < 0.008$
P. trichocarpa TI6	*P. trichocarpa TI7*	*P. trichocarpa TI3*	30/15/68	$\chi^2 = 5.00$, $p < 0.025$
KTI1/2$_{\mathrm{ANC}}$[b]	*KTI6/7*$_{\mathrm{ANC}}$[b]	*KTI3*$_{\mathrm{ANC}}$[b]	25/04/55	$\chi^2 = 15.21$, $p < 0.001$

[a] From Tajima (1993).
[b] Ancestral sequences reconstructed using PAML 3.14.

varsson 2005). A relative rate test (Tajima 1993) comparing rates of sequence evolution of *KTI1* and *KTI2* in *P. tremula*, using the homologous copy from *P. trichocarpa* gwin3 as outgroup, shows significant rate heterogeneity ($\chi^2 = 7.05$, $p < 0.008$; Table 4); *KTI1* has accumulated 28 unique mutations, whereas *KTI2* has only accumulated 11 mutations, since their origin. Similarly, the *KTI6* and *KTI7* paralogues in *P. trichocarpa* also show evidence for rate variation, with *KTI6* and *KTI7* having accumulated 30 and 15 unique mutations, respectively ($\chi^2 = 5.00$, $p = 0.025$; Table 4). Finally, we used PAML to reconstruct ancestral sequences of the *KTI1*/*KTI2* and *KTI6*/*KTI7* clades and compared rates of substitution of the two ancestral paralogous gene copies. The observed rate heterogeneity at the early stage of the duplication of paralogous copies leading to the *KTI1*/*KTI2* and *KTI6*/*KTI7* clades was also significant ($\chi^2 = 15.21$, $p < 0.001$; Table 4).

Discussion

ML analyses confirm previous assertions that *KTI* represents a rapidly evolving gene family in *Populus* (Ingvarsson 2005). Our results show that there is substantial variation in the dN/dS ratio, both between paralogous *KTI* copies and between orthologous copies in different species. A large part of this variation can be attributed to strong diversifying selection resulting from adaptive protein evolution.

The site-specific analysis demonstrates that the strength and direction of natural selection are highly variable among sites in the *KTI* protein. Roughly 35% of codon positions are under strong purifying selection, as evidenced by an ω value close to zero, an additional 40% of the sites appear to be more or less neutral ($\omega = 1$), and the remaining 25% of the sites are under positive selection (models M2 and M8; Table 2). Therefore, the fraction of sites which is identified as having been under positive selection during the diversification of the *KTI* gene family is rather large, with an average $\omega = 2.9$.

These results agree with data on intraspecific nucleotide variation in *Populus tremula*, where several signs of nonneutral evolution have been observed for *KTI1* and *KTI2* (Ingvarsson 2005). Models M2 and M8 identified 13 and 14 sites, respectively, where the Bayesian posterior probability for $\omega > 1$ exceeds 0.95. These sites are primarily located in loop regions connecting different β-strands of the *KTI* protein. At present, it is not clear what functional properties these sites may have. However, it has been speculated that these loop regions are involved in binding the *KTI* protein to the substrate (De Meester et al. 1998). Site-directed mutagenesis at the primary reactive sites of trypsin PIs has been shown to result in a cascade of small structural alterations of the PI molecule that can destabilize the PI-protein complex (Grzesiak et al. 2000). It is, therefore, possible that mutations outside of the reactive loop are crucial for the stabilization of the protease-PI complex. Furthermore, studies of insect PIs have shown that mutations outside the reactive loop can be very important in mediating the specificity of different PIs even if they carry the same residues at the active sites (Kellenberger et al. 2003). In a study of the complete NBS-LLR gene family of pathogen recognition genes in *Arabidopsis thaliana*, Mondragon-Palomino et al. (2002) found that 30% of the positively selected sites were located outside of the LLR regions that mediate direct interactions with pathogen elicitors. While the function of these sites is currently unknown, Mondragon-Palomino et al. (2002) hypothesized that these sites could be important for intermolecular interactions by, for instance, changing the secondary structure of the NBS-LLR proteins.

A priori we expected sites in the reactive loop to be likely targets of positive selection because amino acid differences in the reactive loop are known to result in different binding affinities of PIs to various digestive proteases (De Meester et al. 1998; Song and Suh 1998; Ravichandran et al. 1999). However, none of the sites for which we found strong evidence of positive selection were located in the reactive loop of the *KTI* proteins. Despite this, it is interesting to note that the two residues implicated in mediating the protease inhibitory function (residues P1 and P′ (Song and Suh 1998; De Meester et al. 1998)) are in fact highly variable between, but always conserved within, the different paralogous *KTI* copies (Fig. 2). The pattern of variation at these two residues is worth inspecting in further detail. For instance, *KTI1* and *KTI6* both carry Lys at site 98 (98K); in *KTI1* 97E is paired with Thr at site 97, whereas it is paired with Ile in *KTI6* (97I). Similarly, both *KTI2* and *KTI3* have Glu at site 98 (98E), but *KTI2* has Thr at site 97, whereas *KTI3* has Val. As shown above, our data suggest that the *KTI1*/*KTI2* duplication is fairly recent. Therefore, the presence of either 98K in *KTI1* or 98E in *KTI2* must represent convergent evolution with either *KTI6* or *KTI3*. Such convergent evolution could result from constraints on the number of amino acids that can occur at the inhibitory sites, while still maintaining the protease inhibitory function. Such convergent evolution could perhaps explain why these sites are not identified as targets of positive selection despite differing between paralogous copies. Only model M3 identified site 98 as a likely target of positive selection, with $p > 0.99$; in models M2, M8, and B the Bayesian posterior probability was substantially lower for this site.

The ML analyses that allow for different ω ratios among branches resulting from gene duplication and from speciation events show that newly duplicated genes have diversified under the influence of positive selection ($\omega = 1.86$), whereas following speciation purifying selection appears to be the dominant mode of selection ($\omega = 0.86$; Table 2). This view is also corroborated by the branch-site models (models A and B; Table 2). The latter models also identify three sites as likely targets of positive selection following gene duplication, two of which were not identified in other analyses (Table 2). These sites are also located outside the reactive loop in regions connecting different β-strands, again highlighting the need for further work to understand functional consequences of sites located outside of the reactive loop.

There is evidence for additional variation in dN/dS ratio beyond that explained by the duplication-speciation contrast. An inspection of Fig. 1 suggests that natural selection on the *KTI* genes can be highly episodic even among lineages within a clade. Such episodic selection could be a consequence of the evolution of counter-adaptations in herbivores that lead to an "arms race" between plants and herbivores and thus to an escalation of traits in both plants and herbivores (Bergelson et al. 2001). Newly arisen defense alleles may initially confer a high selective advantage, and if selection is strong enough, such favorable mutations could rapidly increase and fix in a population. However, if the fixation of these newly arisen defense genes is not very rapid, counter-selection in the herbivore population will lessen the effectiveness of the defense functions as the frequencies of these alleles increase in the population. Thus, long periods where several different allelic versions are maintained in the population could be intermixed with occasional rapid fixations of newly arisen mutations (selective sweeps), yielding a very episodic pattern of sequence evolution. Similar results were obtained by Tiffin and Gaut (2001), who found no evidence for natural selection on the serine PI *wip1* in *Zea mays* using intraspecific data. However, significant heterogeneity in the rates of evolution of *wip1* between *Zea mays* and *Sorghum bicolor* led Tiffin and Gaut (2001) to suggest that selection acting on *wip1* was highly episodic.

Relative rate test shows that rates of nucleotide substitutions differ significantly between the two paralogous copies in three independent *KTI* duplications (Table 4). Potentially, such rate heterogeneity could be caused both by variation in the mutation rate between different loci and by natural selection. While we cannot completely rule out mutation rate variation as a cause of rate heterogeneity, divergence at synonymous sites is not significantly different for the different *KTI* genes (Ingvarsson 2005). This suggests that variation in the mutation rate is probably not the cause of the observed rate heterogeneity between duplicate gene copies. Even when rate variation is caused by natural selection, differences can result either from the action of positive selection or from a relaxation of purifying selection, and a demonstration of rate heterogeneity per se therefore does not imply adaptive evolution. However, while the asymmetrical pattern of evolution could be consistent with a period of relaxed selection acting on a duplicate gene copy, results from the codon-based analyses also suggest the action of positive selection.

Gene duplications are known to be important components for the evolution of functional diversification (Ohno 1970; Hughes 1994; Walsh 1995; Lynch and Connery 2000; Moore and Purugganan 2005), and it is easy to envision how functional diversification of *KTI* genes would be favorable in the presence of selection pressures imposed by novel trypsin varieties in rare or new herbivore species. The highly dynamic evolution of the *KTI* genes in *Populus* may therefore represent a counter-strategy to cope with a large number of genes in insects that encode trypsin proteases (Gu et al. 2002; Haq et al. 2004). Similar patterns have been found, for instance, in pathogen recognition genes (*R* genes), which in many plant species have diversified in response to natural selection imposed by numerous pathogen strains (Leister 2004; Meyers et al. 2005). Our data provide strong evidence that Kunitz trypsin inhibitors in *Populus* are a rapidly evolving gene family. Positive Darwinian selection appears to have been driving the diversification of the *KTI* gene family and the selection has been substantially stronger shortly after gene duplication than at other points in the *KTI* phylogeny. Our data thus add to the growing number of studies that have demonstrated rapid evolution in genes involved in coevolutionary processes at the molecular level (e.g., Swanson and Vaquier 2002; Lazzarro et al. 2004; Schein et al. 2004).

Acknowledgments. This work was funded by a grant from the Swedish Research Council (Vetenskapsrådet) to P.K.I. We are grateful to two anonymous reviewers for providing comments that substantially improved the paper.

References

Agrawal S, Tuzun S, Bent E (1999) Inducible plant defenses against pathogens and herbivores: Biochemistry, ecology, and agriculture Am Phytopathol Soc Press, St. Paul, MN

Anisimova M, Bielawski JP, Yang Z (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol 18(8):1585–1592

Bergelson J, Dwyer G, Emerson JJ (2001) Models and data on plant-enemy coevolution. Annu Rev Genet 35:469–499

Bradshaw HD, Hollick JB, Parsons TJ, Clarke HR, Gordon MP (1990) Systemically wound-responsive genes in poplar trees encode proteins similar to sweet potato sporamins and legume Kunitz trypsin inhibitors. Plant Mol Biol 14:51–59

Christopher ME, Miranda M, Major IT, Constabel CP (2004) Gene expression profiling of systemically wound-induced defenses in hybrid poplar. Planta 219:936–947

Clauss MJ, Mitchell-Olds TM (2004) Functional divergence in tandemly duplicated *Arabidopsis thaliana* trypsin inhibitor genes. Genetics 166:1419–1436

Constabel CP (1999) A Survay of Herbivore-Inducible Defensive Proteins and Phytochemicals. In: Agrawal S, Tuzun S, Bent E (eds) Inducible plant defenses against pathogens and herbivores: Biochemistry, ecology, and agriculture. Am Phytopathol Soc Press, St. Paul, MN, pp 137–166

De Meester P, Brick P, Lloyd LF, Blow DM, Onesti S (1998) Structure of the Kunitz-type Soybean trypsin inhibitor (STI): implications for the interactions between members of the STI family and tissue-plasminogen activator. Acta Crystallogr D54:589–597

Grzesiak A, Helland R, Smalas AO, Krowarsch D, Dadlez M, Otlewski J (2000) Substitutions at the P′- position in BPTI strongly affect the association energy with serine proteinases. J Mol Biol 301:205–217

Gu Z, Cavalcanti A, Chen F, Bouman P, Li W (2002) Extent of gene duplication in the genomes of Drosophila, nematode, and yeast. Mol Biol Evol 19:256–262

Haq SK, Atif SM, Khan RH (2004) Protein proteinase inhibitors in combat against insects, pests and pathogens: natural and engineered phytoprotection. Arch Biochem Biophys 431:145–159

Haruta M, Major IT, Christopher ME, Patton JJ, Constable CP (2001) A Kunitz trypsin inhibitor gene family from trembling aspen (*Populus tremuloides* Michx.): cloning, functional expression and induction by wounding and herbivory. Plant Mol Biol 46:347–359

Heitz T, Geoffroy P, Fritig B, Legrand M (1999) The PR-6 family: proteinase inhibitors in plant-microbe and plant-insect interactions. In: Datta SK, Muthukrishan S (eds) Pathogenesis-related proteins in plants. CRC Press, Boca Raton, FL, pp 131–155

Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. Proc R Soc Lond B Biol Sci 256:119–124

Ingvarsson PK (2005) Molecular population genetics of herbivore-induced protease inhibitor genes in European Aspen (*Populus tremula*, L., Salicaceae). Mol Biol Evol 22(9):1802–1812

Karban R, Baldwin IT (1997) Induced responces to herbivory University of Chicago Press, Chicago

Kellenberger C, Ferrat G, Leone P, Darbon H, Roussel A (2003) Selective inhibition of trypsins by insect peptides: Role of P6-P10 loop. Biochemistry 42:13605–13612

Lazzaro BP, Sceurman BK, Clar AG (2004) Genetic basis of natural variation in D-melanogaster antibacterial immunity. Science 303:1873–1876

Lee SI, Lee SH, Koo JC, Chun HJ, Lim CO, Mun JH, Song YH, Cho MJ (1999) Soybean Kunitz trypsin inhibitor (SKTI) confers resistance to the brown planthopper (*Nilaparvata lugens* Stal) in transgenic rice. Mol Breed 5:1–9

Leister D (2004) Tandem and segmental duplication and recombination in the evolution of plant disease resistance genes. Trends Genet 20:116–122

Lynch M, Connery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155

Lynch M, O'Hely M, Walsh B, Force A (2001) The probability of preservation of a newly arisen gene duplicate. Genetics 159:1789–1804

Meyers BC, Kaushik S, Nandety RS (2005) Evolving disease resistance genes. Curr Opin Plant Biol 8:129–134

Michelmore RW, Meyers BC (1998) Clusters of resistance genes evolve by divergent selection and a birth-death process. Genome Res 8:1113–1130

Mondragon-Palomino M, Meyers BC, Michelmore RW, Gaut BS (2002) Patterns of positive selection in the complete NBS-LRR gene family of *Arabidopsis thaliana*. Genome Res 12:1305–1315

Moore RC, Purugganan MD (2003) The early stages of duplicate gene evolution. Proc Natl Acad Sci USA 100:15682–15687

Moore RC, Purugganan MD (2005) The evolutionary dynamics of plant duplicate genes. Curr Opin Plant Biol 8:122–128

Nielsen J, Engelbrecht J, Brunak S, von Heijne G (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Prot Eng 10:1–6

Ohno S (1970) Evolution by gene duplication Springer-Verlag, Berlin

Ravichandran S, Sen U, Chakrabarti C, Dattagupta JK (1999) Cryocrystallography of a Kunitz-type serine protease inhibitor: the 90 K structure of winged bean chymotrypsin inhibitor (WCI) at 2.13 Å resolution. Acta Crystallogr D 55:1814–1821

Ryan CA (1990) Protease inhibitors in plants: genes for improving defenses against insects and pathogens. Annu Rev Phytopath 28:425–449

Saarikoski P, Clapham D, von Arnold S (1996) A wound-inducible gene from *Salix viminalis* coding for a trypsin inhibitor. Plant Mol Biol 31:465–478

Schein M, Yang Z, Mitchell-Olds T, Schmid KJ (2004) Rapid evolution of a pollen-specific oleosin-like gene family from *Arabidopsis thaliana* and closely related species. Mol Biol Evol 21(4):659–669

Song HK, Suh SW (1998) Kunitz-type soybean trypsin inhibitor revisited: Refined structure of its complex with porcine trypsin reveals an insight into the interaction between a homologous inhibitor from *Erythrina caffra* and tissue-type plasminogen activator. J Mol Biol 275:347–363

Swanson WJ, Vaquier V (2002) The rapid evolution of reproductive proteins. Nat Rev Genet 3:137–144

Swanson WJ, Neilsen R, Yang Q (2003) Pervasive adaptive evolution in mammalian fertilization proteins. Mol Biol Evol 20(1):18–20

Tajima F (1993) Simple methods for testing molecular clock hypothesis. Genetics 135:599–607

Tiffin P, Gaut BS (2001) Molecular evolution of the wound-induced serine protease inhibitor *wip1* in *Zea* and related genera. Mol Biol Evol 18:2092–2101

Walsh JB (1995) How often do duplicated genes evolve new functions? Genetics 139:421–428

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15(5):568–573

Yang Z, Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol 46(4):409–418

Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individualnsites along specific lineages. Mol Biol Evol 19(6):908–917

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

Yang Z, Wong WSW, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol 22:1107–1118