Positive selection driving diversification in plant secondary metabolism

Markus Benderoth[†], Susanne Textor[‡], Aaron J. Windsor^{†§}, Thomas Mitchell-Olds^{†§}, Jonathan Gershenzon[‡], and Juergen Kroymann^{†¶}

Department of [†]Genetics and Evolution and [‡]Biochemistry, Max Planck Institute for Chemical Ecology, Hans-Knoell-Strasse 8, D-07745 Jena, Germany

Edited by John F. Doebley, University of Wisconsin, Madison, WI, and approved April 25, 2006 (received for review March 2, 2006)

In *Arabidopsis thaliana* and related plants, glucosinolates are a major component in the blend of secondary metabolites and contribute to resistance against herbivorous insects. Methylthioalkylmalate synthases (MAM) encoded at the MAM gene cluster control an early step in the biosynthesis of glucosinolates and, therefore, are central to the diversification of glucosinolate metabolism. We sequenced bacterial artificial chromosomes containing the MAM cluster from several *Arabidopsis* relatives, conducted enzyme assays with heterologously expressed MAM genes, and analyzed MAM nucleotide variation patterns. Our results show that gene duplication, neofunctionalization, and positive selection provide the mechanism for biochemical adaptation in plant defense. These processes occur repeatedly in the history of the MAM gene family, indicating their fundamental importance for the evolution of plant metabolic diversity both within and among species.

biochemical neofunctionalization | glucosinolate metabolism | methylthioalkylmalate synthase | plant–enemy interactions

Plants synthesize an immense number of secondary compounds, so called because their significance for processes of basic growth and development is not immediately evident. More than 200,000 known secondary metabolites provide an increasingly exploited reservoir for the generation of pharmaceutically active agents (1), and many more await discovery. Classic hypotheses that seek to explain this vast metabolic diversity propose a stepwise and reciprocal process of adaptation and counteradaptation between plants and their natural enemies, molded by mutual selection (2). In *Arabidopsis thaliana* and other crucifers, glucosinolates are a major component in the mélange of secondary metabolites. More than 120 glucosinolates are known, which share a chemical core structure but differ in their amino acid-derived side chain. Glucosinolate composition and quantity varies among and within species (3–5). Upon tissue disruption, myrosinase-catalyzed hydrolysis of glucosinolates generates biologically active compounds, which play an important ecological role in plant defense against herbivorous insects (4, 6–9). However, insects can adapt to glucosinolate profiles or evade deleterious effects from glucosinolate hydrolysis by counteradaptation (9–12). In *A. thaliana*, a modular genetic system regulates diversity in glucosinolate profiles among accessions and may permit the rapid generation of new glucosinolate combinations in response to challenges imposed by the biotic environment (3). Several genetic loci (3, 9), each present in several alleles, interact epistatically and, together with modifying proteins (7), determine the blend of bioactive products that emerges during glucosinolate hydrolysis (9). One genetic locus central for glucosinolate diversity, the methylthioalkylmalate synthase gene (*MAM*) cluster, controls an early step in the biosynthesis of aliphatic glucosinolates, the predominant glucosinolate class in *A. thaliana*. This locus comprises a small family of *MAM* genes, encoding methylthioalkylmalate synthases that condense ω -methylthio-2-oxoalkanoic acids derived from methionine with acetyl-CoA in a reaction series that extends the ω -methylthio-2-oxoalkanoic acid by one methylene group with each reaction cycle (13) (Fig. 5, which is published as supporting information on the PNAS web site). Thereby, they provide a diverse

mixture of precursors for subsequent biosynthetic reaction steps. In *A. thaliana*, the archetypical configuration of the *MAM* cluster comprises three genes, *MAM2*, *MAM1*, and *MAM-L*. However, this composition varies among accessions because of gene deletion and conversion events, causing a quantitative trait locus for resistance against herbivorous insects (4). *MAM-L* provides precursors for aliphatic glucosinolates with long side chains (14), whereas *MAM2* and *MAM1* control short-chain glucosinolates (4, 13, 15).

Here, we conduct MAM enzyme assays and analyze *MAM* sequence variation patterns in a phylogenetic framework to further elucidate the evolutionary forces that lead to functional divergence in the *MAM* gene cluster.

Results

Comparative Sequencing and Phylogenetic Analyses Reveal Complex Relationships of MAM Genes Among Arabidopsis Relatives. We sequenced the *MAM* region from several close *A. thaliana* relatives, *Arabidopsis lyrata* ssp. *lyrata*, *Arabidopsis lyrata* ssp. *petraea*, and *Arabidopsis ceibensis*, and the more distantly related *Boechera divaricarpa* (16). We found that the region is colinear between *A. thaliana* and its relatives (Fig. 1), consistent with other studies demonstrating extensive synteny between cruciferous genomes (17–20). In *A. lyrata*, *A. petraea*, *A. ceibensis*, and *B. divaricarpa*, the *MAM* cluster consists of three genes, which we designate *MAMa*, *MAMb*, and *MAMc*. Phylogenetic analyses (Fig. 2; see Fig. 6, which is published as supporting information on the PNAS web site) show that *A. thaliana* *MAM1* and *MAM2* originate from a *MAMa* gene duplication after *A. thaliana* diverged from its congeners. Several estimates of evolutionary rates in the Brassicaceae are in rough agreement (16, 21). From these rates, we concluded that *MAM1* and *MAM2* have existed in *A. thaliana* for $>10^5$ generations. Duplication events leading to tandemly linked *MAMa*, *MAMb*, and *MAMc* are much older and predate the divergence of *Arabidopsis* and *Boechera* genera >10 million years ago (16). *A. thaliana* *MAM-L* is orthologous to *MAMb*, and *MAMc* has been lost in *A. thaliana*.

Enzyme Assays Reveal MAM1 Biochemical Neofunctionalization. The *MAM1* gene product from the *A. thaliana* accession Columbia (Col-0) catalyzes the condensation of both 4-methylthio-2-oxobutanoic acid (13) and 5-methylthio-2-oxopentanoic acid (15) with acetyl-CoA (Fig. 5). Thus, Col-0 *MAM1* controls the committed step in both the first and second round of methionine carbon chain elongation, ultimately leading to the formation of aliphatic glucosinolates with four methylene groups (4C) in their side chain. We expressed *MAM2* from the *A. thaliana* accession Landsberg

Conflict of interest statement: No conflicts declared.

This paper was submitted directly (Track II) to the PNAS office.

Abbreviation: MAM, methylthioalkylmalate synthase.

Data deposition: The sequences reported in this paper have been deposited in the EMBL database (accession nos. AM180569–AM180585).

[§]Present address: Department of Biology, Duke University, Durham, NC 27708.

[¶]To whom correspondence should be addressed. E-mail: kroymann@ice.mpg.de.

© 2006 by The National Academy of Sciences of the USA

Fig. 2. Phylogenetic relationships between *MAM* genes, according to the neighbor-joining method. The significance of the branching order was tested by bootstrapping (2,000 replicates). Isopropylmalate synthase genes from *A. thaliana* (At1g18500 and At1g74040; ref. 13) and rice (GenBank accession no. AK066890) were used as references. Note that branches indicated by an asterisk were shortened to half of its original lengths. Bold lines indicate branches with evidence for positive selection in maximum likelihood analyses.

ymous changes along the branches of a tree connecting *A. lyrata* *MAMa* with *A. thaliana* *MAM1* and *MAM2* and tested ratios with a heterogeneity G test. We found heterogeneous ratios ($\chi^2 = 10.76$; $df = 2$; $P = 0.0046$) because of an excess of replacement changes on branch b (Fig. 3), leading from the *MAMa* duplication to *MAM1*. Finally, we analyzed codon substitution patterns with a maximum likelihood approach implementing a branch-site model (23–25). Models that allowed for positive selection acting on *MAM1* subsequent to the *MAMa* duplication fit the data significantly better than a nearly neutral model that allowed for neutral or relaxed but not positive selection (Table 2; see also Table 3, which is published as supporting information on the PNAS web site). Moreover, PAML indicated positive selection for most amino acids that contrast *MAM1* from *MAM2* or *MAMa* (Fig. 4 and Table 3).

In all investigated species, the *MAM* cluster consists of several gene copies, reflecting ancestral gene duplication events. Therefore, we also tested a data set comprising *MAMa*, *MAMb*, and *MAMc* sequences from our species survey for potential positive

Table 1. McDonald–Kreitman tests for selection

Plant	Complete coding	Residue exclusion
<i>A. lyrata</i> <i>MAMa</i>	6.999 (<0.01)	9.239 (<0.01)
<i>B. divaricarpa</i> <i>MAMa</i>	4.245 (<0.05)	5.547 (<0.05)

Statistical tests for Sorbo *MAM1* and *MAM2* replacement to synonymous nucleotide ratios by using *A. lyrata* or *B. divaricarpa* *MAMa* as references. Calculations change of G values are either based on complete coding sequences or exclude residues corresponding to presumptive chloroplast targeting sequences. We also tested different combinations of *MAM1* and *MAM2* sequences from a previous sequence survey (4) and found that results were insensitive to the combination of *MAM1* and *MAM2* sequences included in the analyses.

Fig. 3. Gene duplication and speciation. Based on parsimony, we partitioned nonsynonymous (N) and synonymous changes (S) along branches of a simplified phylogeny connecting *A. lyrata* *MAMa*, *A. thaliana* Sorbo and Col-0 *MAM1*, and Sorbo and Ler-0 *MAM2*. We excluded two synonymous sites with complex substitution patterns from the analyses. Ratios of nonsynonymous versus synonymous changes are heterogeneous across the tree ($\chi^2 = 10.76$; $df = 2$; $P = 0.0046$) because of an excess of replacement changes on branch b.

selection. Again, models that allowed for positive selection subsequent to gene duplication events (i.e., along basal branches) explained the data significantly better than models that incorporated only neutral evolution or models that allowed for positive selection only on distal branches of the tree (Table 2; see also Table 4, which is published as supporting information on the PNAS web site).

Discussion

We provide evidence that *MAM1* and *MAM2* originate from a duplication of *MAMa* and that *MAM1*, but not *MAM2*, has been subject to positive selection. We surveyed the enzymes' biochemical properties, e.g., pH-, ATP-, and metal ion dependency (Table 5 and Fig. 8, which are published as supporting information on the PNAS web site) and found no major difference except for the disparity in substrate acceptance between *MAMa*, *MAM2* (both acting only on 2C substrates), and *MAM1* (accepting 2C and 3C substrates and minor amounts of 4C substrate). Our *in vitro* assays are supported by *in planta* data; accessions like Col-0 or Sorbo with a functional *MAM1* gene accumulate aliphatic glucosinolates derived from dihomomethionine (4C), whereas accessions such as Ler-0 that lack a functional *MAM1* (Fig. 1) contain predominantly homomethionine-derived (3C) aliphatic glucosinolates (4, 13). Therefore, we conclude that the additional enzymatic capacity of *MAM1* was the target of positive selection.

What are the consequences of this biochemical neofunctionalization? Enzymes encoded at the *MAM* locus are central to the

Table 2. Analysis of codon substitution patterns

Test	Model	lnL	$P_{df=2}$
Test for positive selection on <i>MAM1</i>	M1a: NearlyNeutral	−3547.83	
	Branch-site Model A	−3541.87	<0.01
	Branch-site Model A ($\omega = 1$)	−3543.66	<0.05
Test for positive selection along basal branches connecting <i>MAMa</i> , <i>MAMb</i> , and <i>MAMc</i>	M1a: NearlyNeutral	−5086.90	
	Branch-site Model A	−5070.99	<0.001
	Branch-site Model A ($\omega = 1$)	−5074.02	<0.001

PAML 3.14 was used to compute log likelihoods (lnL) for null models M1a and branch-site models A. Two different subsets of *MAM* sequences were tested for positive selection acting on *MAM1* or along basal branches of the *MAM* phylogeny (see *Materials and Methods*). In the examples shown in the table, codon frequencies were calculated from the average nucleotide frequencies at the three codon positions. The branch-site models explain the data significantly better than the nearly-neutral models, and statistical significance (P) is indicated. Complete results from our analyses are shown in Tables 3 and 4.

gen) according to the manufacturer's instructions. We amplified *MAM* cDNAs with AccuPrime Taq polymerase (Invitrogen) by using the following primer pairs: *Ler-0 MAM2*, 5'-ATGTCA-TRTTGCTCTTCTGTG-3' and 5'-CACATTAGATGAAAC-CTGAA-3'; *Sorbo MAM1*, 5'-ATGTCATRTTGCTCTTCTGTG-3' and 5'-CACATTGCGAYGAAACCTGA-3'; *A. lyrata MAMa*, 5'-ATGTCATGTTGCTCTTCTGTG-3' and 5'-TAGCATATTTGATGAAACCT-3'; *B. stricta MAMa*, 5'-ATGTCATGTTGCTCTTCTGA-3' and 5'-GAGGGAAAC-CTGAGGAAC-3'. We chose primer pairs such that expressed gene products lacked N-terminal signal sequence (13) but contained a C-terminal His tag extension encoded by the pCR T7/CT TOPO (Invitrogen) expression vector. PCR consisted of an initial denaturation at 94°C for 1 min, and 33 cycles at 94°C for 15 s, 55°C for 40 s, and 72°C for 2 min, followed by a final elongation at 72°C for 6 min on a PE Applied Biosystems 9700 thermal cycler. We separated PCR products on 1% agarose gels, excised them, and gel-purified by using the NucleoSpin Extract Kit (Macherey-Nagel) according to the manufacturer's instructions. We cloned purified cDNA into pCR T7/CT TOPO, and electrotransformed into *E. coli* TOP-10 cells (Invitrogen). For each *MAM* cDNA, we isolated and sequenced several plasmid clones to confirm sequence integrity, reading frame, and direction within the vector.

To test which portion of the *A. thaliana* sequences determined substrate specificity of *MAM1* versus *MAM2*, we excised nucleotides 163–1031 from a functional *Sorbo MAM1* expression construct with BamHI/BsmI and ligated it with the vector-containing portion of an identically treated functional *Ler-0 MAM2* construct and vice versa. Thus, we obtained two fusion constructs, one with partial *Sorbo MAM1* cloned into *Ler-0 MAM2* sequence and the second with partial *Ler-0 MAM2* cloned into *Sorbo MAM1* sequence.

Heterologous Expression. We transformed *MAM* constructs into *E. coli* BL21 (DE3) (Invitrogen) heterologously in Overnight-Express Instant TB Medium (Novagen) according to the manufacturer's instructions. We isolated total soluble protein with the BugBuster/Lysosome system (Novagen), and we determined protein concentration with the D_c Protein Assay (Bio-Rad). We confirmed presence, size, and expression intensity of the target protein with SDS/PAGE and Western blot analysis.

Enzyme Assays. Enzyme assays followed the basic protocol described in ref. 13 and modified according to ref. 15. To determine pH, ATP, and metal ion dependencies, we carried out MAM activity assays with 80 mM Tris-HCl (pH 8.0)/3 mM 4-methylthio-2-oxobutanoic acid (the 2C substrate, see Fig. 5)/1 mM acetyl-CoA/4 mM Mn²⁺ as initial conditions. These conditions work well with *A. thaliana* Col-0 *MAM1* (13, 15). Then, we measured pH dependency between pH 6.5 and 9.0 in incremental steps of 0.25 (Table 5). For subsequent assays, we used the pH optimum determined for each single MAM synthase. Next, we optimized ATP concentration by measuring ATP dependency between 0.0 and 6.0 mM in incremental steps of 0.5 mM. Again, in subsequent assays, we chose ATP optima determined for each single MAM synthase. Finally, we measured Mn²⁺ dependency in concentrations between 0.0 and 9.6 mM in incremental steps of 0.8 mM. We replicated each assay at least three different times. We then used these optimal assay conditions for each MAM synthase to determine activity with the native 2C, 3C, and 4C substrates (4-methylthio-2-oxobutanoic acid, 5-methylthio-2-oxopentanoic acid, and 6-methylthio-2-hexanoic acid, respectively; see Figs. 5 and 8) and with the artificial 2C, 3C, and 4C substrates (2-oxohexanoic acid, 2-oxoheptanoic acid, and 2-oxooctanoic acid, respectively; see Fig. 8). We used the artificial substrates, with the thiol group in the side chain replaced by a methylene function, to confirm the pattern of *in vitro* substrate specificity revealed by the use of the natural

substrates. These analogues were shown previously to be accepted by *A. thaliana* Col-0 *MAM1* at efficiencies slightly less than those of the native substrates (15). In addition, we measured MAM activity with different metal anions (Ca²⁺, Co²⁺, Cu²⁺, and Mg²⁺; all added as Me²⁺Cl₂) by using MAM synthase-specific pH and ATP optima and compared it to MAM activity with Mn²⁺ (Table 5). In these assays, all metal anions were added to a final assay concentration of 5.6 mM. Again, each experiment was repeated at least three times. Because of the large number of replicated enzyme assays and because we found no difference in substrate specificity between crude protein extracts and purified proteins, we mostly used crude protein extracts from heterologous expression for our analyses, with a typical assay containing between 20% and 30% target protein in 100 μg of crude protein.

Enzyme Purification. We purified *MAM* proteins by using the His-Bind column and buffer system (Novagen). We used Vivaspins 15R centrifuge concentrator units with a molecular weight cut off of 10,000 (Sartorius) for rebuffing and concentrating purified protein.

Evolutionary Analyses. We used CLUSTALX 1.8 to align *MAM* protein sequences and corrected gap locations manually. We removed alignment positions corresponding to gaps comprising >1 aa in any sequence and converted the amino acid alignment to a nucleotide alignment. We used this data set for all phylogenetic analyses. We used TREECON (39) for the construction of a neighbor-joining tree, PAUP 4.0b10 (40) for maximum likelihood estimation (41), and MRBAYES 3.1 (42) for Bayesian inference (43). For both maximum likelihood and Bayesian inference, we assumed six substitution types and based our calculations on empirical nucleotide frequencies. We approximated rate distribution at variable sites by a gamma distribution with four rate categories and an estimated shape parameter α . These settings correspond to the general time reversible with gamma estimation model. We obtained the maximum likelihood seed tree by stepwise addition and used tree-bisection-reconnection as the branch swapping algorithm. We performed Bayesian analysis by simultaneous calculation of four Markov chain Monte Carlo runs, each with four chains of 5 million iterations. MRBAYES default settings were used to calculate prior probabilities. Irrespective of whether we used distance matrix, maximum likelihood, or Bayesian inference methods, we obtained the same tree branching order.

DNASP 4.00 (44) was used to calculate replacement and synonymous nucleotide polymorphisms.

Analysis of Codon Substitution Patterns. We used PAML3.14 (24) to analyze codon substitution patterns with a maximum likelihood approach, implementing a branch-site model (Model A) (23–25). To test for potential positive selection subsequent to the *MAMa* gene duplication, we used two different alignment files, “*MAM1-MAM2-MAMa*” and “*MAM1-MAM2-MAMa trunc*” (Table 3). The first file includes all *MAM* nucleotide positions, whereas the second alignment file excludes nucleotides corresponding to *MAM* chloroplast targeting signals. Likewise, we used two different alignment files, “*MAMa-MAMb-MAMc*” and “*MAMa-MAMb-MAMc trunc*” to test for positive selection along ancestral phylogenetic branches (Table 4). These alignments are slightly shorter (incorporating alignment positions 1–1461 or 142–1461, respectively) because of a mutation in *Boechera divaricarpa* that leads to a *MAMc* protein with a truncated C terminus. In Table 3, “*ColMAM1*” and “*SorboMAM1*” refer to *MAM1* sequences from *A. thaliana* accessions Col-0 and Sorbo, respectively. “*LerMAM2*,” “*SorboMAM2*,” and “*TsuMAM2*” refer to *MAM2* sequences from *A. thaliana* accessions Ler-0, Sorbo, and Tsu-0. “*AL-MAMa*,” “*APMAMa*,” “*ACMAMa*,” and “*BDMAMa*” refer