Journal of Biochemistry and Molecular Biology, Vol. 40, No. 2, March 2007, pp. 247-260

JOURNAL OF Biochemistry and Molecular Biology

Molecular Cloning of Two Genes Encoding Cinnamate 4-Hydroxylase (C4H) from Oilseed Rape (*Brassica napus*)

An-He Chen[1,2,†], You-Rong Chai[1,†], Jia-Na Li[1,†,*] and Li Chen[1]

[1]Chongqing Rapeseed Technology Research Center; Chongqing Key Laboratory of Crop Quality Improvement; Key Lab of Biotechnology & Crop Quality Improvement of Ministry of Agriculture; College of Agronomy and Life Sciences, Southwest University, Beibei, Chongqing 400716, People's Republic of China

[2]College of Bio-Information, Chongqing University of Posts and Telecommunications, Huang jueya, Nanan Zone, Chongqing 400065, People's Republic of China

Received 19 April 2006, Accepted 27 November 2006

Cinnamate 4-hydroxylase (C4H) is a key enzyme of phenylpropanoid pathway, which synthesizes numerous secondary metabolites to participate in development and adaption. Two *C4H* isoforms, the 2192-bp *BnC4H-1* and 2108-bp *BnC4H-2*, were cloned from oilseed rape (*Brassica napus*). They both have two introns and a 1518-bp open reading frame encoding a 505-amino-acid polypeptide. BnC4H-1 is 57.73 kDa with an isoelectric point of 9.11, while 57.75 kDa and 9.13 for BnC4H-2. They share only 80.6% identities on nucleotide level but 96.6% identities and 98.4% positives on protein level. Showing highest homologies to *Arabidopsis thaliana* C4H, they possess a conserved p450 domain and all P450-featured motifs, and are identical to typical C4Hs at substrate-recognition sites and active site residues. They are most probably associated with endoplasmic reticulum by one or both of the N- and C-terminal transmembrane helices. Phosphorylation may be a necessary post-translational modification. Their secondary structures are dominated by alpha helices and random coils. Most helices locate in the central region, while extended strands mainly distribute before and after this region. Southern blot indicated about 9 or more *C4H* paralogs in *B. napus*. In hypocotyl, cotyledon, stem, flower, bud, young- and middle-stage seed, they are co-dominantly expressed. In root and old seed, *BnC4H-2* is dominant over *BnC4H-1*, with a reverse trend in leaf and pericarp. Paralogous *C4H* numbers in Brassicaceae genomes and possible roles of conserved motifs in 5' UTR and the 2nd intron are discussed.

Keywords: Cinnamate 4-hydroxylase, Cloning, Expression, Oilseed rape (*Brassica napus*)

Introduction

Phenylpropanoid pathway produces a large number of biologically important secondary metabolites through several important branch pathways. One of them synthesizes lignins, which play fundamental roles in mechanical support, solute conductance and disease resistance in higher plants (Barber and Mitchell, 1997; Harakava, 2005). Another important branch pathway synthesizes various flavonoid compounds. In addition to attracting pollinators and protecting plants from UV irradiation and attacks by fungi and animals, flavonoids also possess anti-inflammatory, antiallergenic, antioxidant or cancer preventive functions in human (Benavente-Garcia *et al.*, 1997; Di Carlo *et al.*, 1999; Harborne and Williams, 2000; Manthey *et al.*, 2001; Le Marchand, 2002). Phenylpropanoid pathway also synthesizes coumarins, salicylic acid, isoflavonoids (phytoalexins), chlorogenic acids and stilbenes to act as

Database Accession Nos: **DQ485129** and **DQ485130** for *BnC4H-1* gene and mRNA, and **DQ485131** and **DQ485132** for *BnC4H-2* gene and mRNA, respectively.

Abbreviations: bp, base pair; C3'H, p-coumaroyl CoA shikimate/quinate 3'-hydroxylase; C4H, cinnamate 4-hydroxylase; cDNA, complementary DNA; CTAB, cetyl trimethyl ammonium bromide; DAF, d after flowering; F3'H, flavonoid 3'-hydroxylase; F3'5'H, flavonoid 3',5'-hydroxylase; F5H, ferulate-5-hydroxylase; ORF, open reading frame; P450, cytochrome P450; PCR, polymerase chain reaction; pI, isoelectric point; RACE, rapid amplification of cDNA ends; RT, reverse transcription; (S)-N-M3'H, (S)-N-methylcoclaurine 3'-hydroxylase; UTR, untranslated region.

†These authors have equal contribution to the study.

*To whom correspondence should be addressed.
Tel: 86-23-68251950; Fax: +86-23-68251950
E-mails: chaiyourong@tom.com or ljn1950@swu.edu.cn

signaling molecules or antagonistic ingredients (Dixon and Paiva, 1995; Dixon *et al.*, 1996; Weisshaar and Jenkins, 1998; Dixon and Steele, 1999). Manipulation of phenylpropanoid pathway metabolites has long been a hotspot (Dixon *et al.*, 1996; Dixon and Steele, 1999).

Cinnamate 4-hydroxylase (C4H, **EC 1.14.13.11**) catalyzes the hydroxylation of *trans*-cinnamic acid to 4-hydroxycinnamate and is the second key enzyme of common phenylpropanoid pathway (Russell, 1971). It belongs to CYP73A subfamily of cytochrome P450-dependent monooxygenase superfamily. Core enzymes of phenylpropanoid metabolism are believed to form an enzyme complex, and C4H plays a pivotal role at the interface between cytosolic phenylpropanoid pathway and membrane-localized electron-transfer reactions (Chapple, 1998; Koopmann *et al.*, 1999; Winkel-Shirley, 1999). C4H was first purified in 1991 (Gabriac *et al.*, 1991). According to Dr. Nelson's P450s database (**http://drnelson.utmem.edu/biblioD.html#73A**), 54 *C4H* genes have been isolated from various plant species such as alfalfa, *Arabidopsis*, artichoke, etc. (Fahrendorf and Dixon, 1993; Mizutani *et al.*, 1993; Teutsch *et al.*, 1993).

Oilseed rape (*Brassica napus* L.) is one of the five major oil crops in the world. In this crop, many agronomically important traits related to phenylpropanoid pathway are focuses of genetic improvement pursued by researchers for many years. For example, the commonly occurred lodging problem calls for stronger stems and branches. Improvement of resistance to diseases needs quicker and enhanced cell wall lignification in response to pathogen invasion. Genetic engineering of lignin pathway flux, monolignol ratio and lignin composition provides a promising strategy to cope with these problems (Anterola and Lewis, 2002). In recent years yellow seed trait of *B. napus* has attracted many researchers due to its good quality. However, lacking of yellow-seeded genotypes together with instability of yellow seed phenotype has largely retarded breeding and application of yellow-seeded rapeseed (Heneen and Brismar, 2001). The mechanism of yellow seed trait formation of *B. napus* is still not clear. The most typical feature of yellow seed trait is the reduction of lignin and pigment contents in the seed coat. As has been revealed, plant seed coat pigments are polymers of proanthocyanidin, a metabolite of flavonoid pathway (Debeaujon *et al.*, 2003). Study on *B. napus C4H* gene will help dissect the mechanism of yellow seed trait formation and lay the base for transgenic creation of stable yellow-seeded *B. napus*.

In family Brassicaceae, except the characterized *C4H* gene from *Arabidopsis thaliana*, no other full-length *C4H* gene has been cloned, though many important oilseed and vegetable crops are included in this family. Here we report the cloning and molecular characterization of two isoforms, *BnC4H-1* and *BnC4H-2*, of *C4H* gene family from *B. napus*. Our work enables further investigation of the roles *C4H* genes play in determining many important traits, and will undoubtedly provide the possibility to improve disease resistance and anti-lodging ability, as well as to create artificial yellow seed trait of oilseed rape through regulating the expression levels of *C4H* genes.

Materials and Methods

Vectors and strains. *Escherichia coli* strain DH5α originally offered by Professor Kexuan Tang, School of Life Sciences, Fudan University was preserved by our laboratory. T-vector pMD18-T was the product of Takara Biotechnology (Dalian) Co., Ltd.

Plant materials. The plant materials used here including root, hypocotyl, cotyledon, stem, leaf, bud, flower, silique pericarp, and seed of 10, 20 and 30 d after flowering (DAF) of *B. napus* stock line 5B were sampled from the experimental field of Southwest University, China. The samples were immediately frozen in liquid nitrogen, and preserved at −80°C.

RNA and DNA isolation. Total RNA of each tissue sample was extracted using a CTAB method described by Jaakola (Jaakola *et al.*, 2001). All RNA samples were digested with RNase-free DNase I (Worthington) to remove contaminated DNA. Total genomic DNA was isolated using a CTAB-based method (Rechards, 1995). The quality and concentration of RNA and DNA samples were examined by agarose gel electrophoresis and spectrophotometer analysis.

3' and 5' cDNA end amplification of *C4H* genes from *B. napus*. An aliquot of 5 μg equally proportioned (w/w) mixture of total RNA from various organs was used as template to generate first strand total cDNA using GeneRacer Kit (Invitrogen) in terms of manual instruction. Based on multi-alignment (Vector NTI Advance 9.0) of *C4Hs* from *A. thaliana* and other plants, forward primers FC4H3-1 (5'-TGATGATGTACAACAACATGTTCCG-3') and FC4H3-2 (5'-CCTCACATGAACCTCCATGATGC-3') corresponding to two conserved sites were synthesized for 3' rapid amplification of cDNA ends (RACE) of *B. napus C4H* genes. FC4H3-1 was paired with GeneRacer 3'-Primer to carry out the primary amplification of 3' RACE in a standard 50-μl *Taq* PCR system containing 0.5 μl total cDNA as template. Amplification conditions were as follows: predenaturation at 94°C for 2 min, followed by 25 cycles of amplification (94°C for 1 min, 50°C for 1 min, 72°C for 1 min 30 s) and by 72°C for 10 min. One μl of 50-fold diluted PCR product was used as template for 3'-nested PCR using primer FC4H3-2 and GeneRacer 3'-Nested Primer with an anneal temperature of 55°C. After agarose gel electrophoresis, DNA of the target band was recovered (Gel Extraction Mini Kit, Watson Biotechnologies, Inc.) and ligated to pMD18-T for transformation of DH5α via a $CaCl_2$ method (Seidman *et al.*, 1995). Positive colonies were sequenced using primers M13F/M13R at Shanghai Bioasia Company, China.

In 5' RACE, antisense primer RC4H5-1 (5'-GCATCATGGAGG TTCATGTGAGG-3') was synthesized to pair with GeneRacer 5'-Primer to conduct the primary amplification. While antisense primer RC4H5-2 (5'-CGGAACATGTTGTTGTACATCATCA-3') paired with GeneRacer 5'-Nested Primer was adopted for nested PCR. The cycling conditions were the same as those for 3' RACE primary PCR. Gel recovery, TA cloning and sequencing were performed.

Amplification of full-length cDNAs and genomic sequences of *B. napus C4H* genes. Based on sequencing results of the 3' and 5' RACE products, sense primers FBNC4-2 (5'-AGCAGCTCCTTCT GCTTTCTC-3') and FBNC4-3 (5'-TCAGCAGCTCCTTCTGCTT TC-3'), and antisense primers RBNC4-1 (5'-CAAAACAGTGGGA ACCAATAGTTATTG-3') and RBNC4-7 (5'-CCGAAGAAACAA CACATTGAATATTCAAC-3'), were designed corresponding to the 5' and 3' cDNA ends. They were combined into 4 primer pairs, i.e. FBNC4-2/RBNC4-1, FBNC4-2/RBNC4-7, FBNC4-3/RBNC4-1 and FBNC4-3/RBNC4-7, for amplification of full-length cDNAs. The 50-μl standard *Taq* PCR system and following cycling parameters were used: 94°C for 3 min, followed by 30 cycles of amplification (94°C for 1 min, 62°C for 1 min, 72°C for 2 min 30 s) and by 72°C for 10 min. Corresponding genomic sequences were amplified by replacing the template with 0.5 μg total genomic DNA under the same conditions. Gel recovery, TA cloning and sequencing were performed.

Sequence alignment, open reading frame (ORF) translation and molecular weight calculation of predicted proteins were carried out with Vector NTI Advance 9.0. BLAST was done at the NCBI server (**http://www.ncbi.nlm.nih.gov/BLAST/**), while structural analysis of deduced proteins was carried out on the website (**http://cn.expasy.org/tools/**).

Southern blot analysis. Southern blotting was carried out to analyze the copy numbers of *C4H* genes in the *B. napus* genome. Forty-μg aliquots of total genomic DNA were digested overnight at 37°C with *Dra*I, *Eco*RI, *Eco*RV and *Hin*dIII (MBI Fermentas), which did not cut within the probe region, respectively, fractionated by 0.80%-agarose gel electrophoresis, and transferred to a positively charged nylon membrane (Roche) through capillarity (Sambrook and Russell, 2001). Primer pair FBNC4-2/RC4H5-2 was used to amplify a 656-bp conserved fragment using BnC4H-1 full-length cDNA as template. PCR DIG Probe Synthesis Kit (Roche) was used to label the probe with Digoxigenin-11-dUTP under the following procedure: 95°C for 2 min, followed by 30 cycles of amplification (95°C for 30 s, 60°C for 30 s, 72°C for 40 s), succeeded by 10 min at 72°C. Hybridization was carried out at 42°C for 16 h (DIG Easy Hyb, Roche). After stringent washing and immunological detection with the DIG Wash and Block Buffer Set and DIG Nucleic Acid Detection Kit (Roche), the hybridization bands were pictured.

RT-PCR detection of transcripts of *BnC4H-1* and *BnC4H-2* in various organs of *B. napus*. Semi-quantitative RT-PCR was performed to detect the transcription levels of *BnC4H-1* and *BnC4H-2* in 11 organs of *B. napus*. Oligo $(dT)_{20}$-directed reverse transcription of 5-μg total RNA of each sample was performed using SuperScript III First-Strand Synthesis SuperMix (Invitrogen, USA). Primers FBNC4-2 and RBNC4-2N (5'-TTTGGTGAGGTT CGGGGAG-3') were used to isoform-specifically amplify a 357-bp region of *BnC4H-1*, while FBNC4-1 and RBNC4-1N (5'-CCTTTC GTGGCCGAATCAAG-3') for specific amplification of a 516-bp region of *BnC4H-2*. An aliquot of 0.5-μg first strand cDNA of each sample was taken as template in a 50-μl standard *Taq* PCR reaction. The cycling procedure was: 94°C for 2 min, followed by 30 cycles of amplification (94°C for 1 min, 62°C for 1 min, 72°C for 1 min), then 72°C for 10 min. To identify the uniformity of total first strand cDNA among samples, *Arabidopsis*-based primers FATACT2 (5'-GTGTTGTTGGTAGGCCAAGACATCA-3') and RATACT2 (5'-CTTGATGTCTCTTACAATTTCCCGC-3') were designed to amplify the *actin* gene fragment orthologous to a 542-bp region of *A. thaliana ACT2* under the same conditions annealed at 55°C. PCR products were detected by agarose gel electrophoresis and pictured.

Results

Sequence cloning of *BnC4H-1* and *BnC4H-2*

3' RACE result. Agarose gel electrophoresis revealed that amplification with primers FC4H3-1 and GeneRacer 3' Primer resulted in 2 bands, one about 1000 bp and the other about 1200 bp, accompanied by some faint bands and smear. Nested amplification of the 3' RACE resulted in a bright band of about 600 bp, also accompanied by 2 faint bands (400 bp and 700 bp) and some smear. The 600-bp band was in consensus with homology-based prediction, so it was recovered and subcloned. Eight clones were sequenced and two different 3' cDNA ends were obtained. One was 496 bp and the other was 524 bp, not including the poly(A) tail. NCBI blastn analysis indicated that they showed wide homologies to known *C4H* sequences, with the highest identities to *A. thaliana C4H* mRNA (**NM_128601**).

5' RACE result. Gel detection of primary PCR product of 5' RACE showed a band of about 1200 bp and 3 faint bands of about 1100, 700 and 250 bp. Nested amplification yielded a specific bright band of about 700 bp, which was also in consensus with homology-based length prediction. Five sequenced clones resulted in 2 different 5' cDNA ends. One was 656 bp and the other was 658 bp (the 30-bp GeneRacer RNA Oligo-derived sequence removed). Their highest similarities to known *C4H* sequences were proved by blastn.

Amplification of full-length cDNAs and genomic sequences of *B. napus C4H* genes. The 4 primer pairs all yielded specific bright bands of about 1750 bp in full-length cDNA amplifications, but the bands of FBNC4-2/RBNC4-1 and FBNC4-3/RBNC4-1 were a little longer than those of FBNC4-2/RBNC4-7 and FBNC4-3/RBNC4-7. All the 4 bands were recovered, subcloned and sequenced. Sequencing results of the 2 longer bands were identical to each other except for the 2-bp difference caused by sense primers, and the same case with the 2 shorter bands. This indicated that only 2 full-length cDNAs, denoted *BnC4H-1* and *BnC4H-2* here, were obtained practically. They were identical to the two 5' and the two 3' cDNA ends in corresponding regions, and alignment indicated that the right primer combinations for *BnC4H-1* and *BnC4H-2* were FBNC4-2/RBNC4-1 and FBNC4-3/RBNC4-7 respectively. So these 2 primer pairs were used to amplify the genomic sequences of *BnC4H-1* and *BnC4H-2*. Gel detection showed an about 2200-bp band for FBNC4-2/RBNC4-1 and an about 2100-bp band for FBNC4-3/RBNC4-7. Sequencing results of them were identical to the corresponding cDNAs except the intron regions.

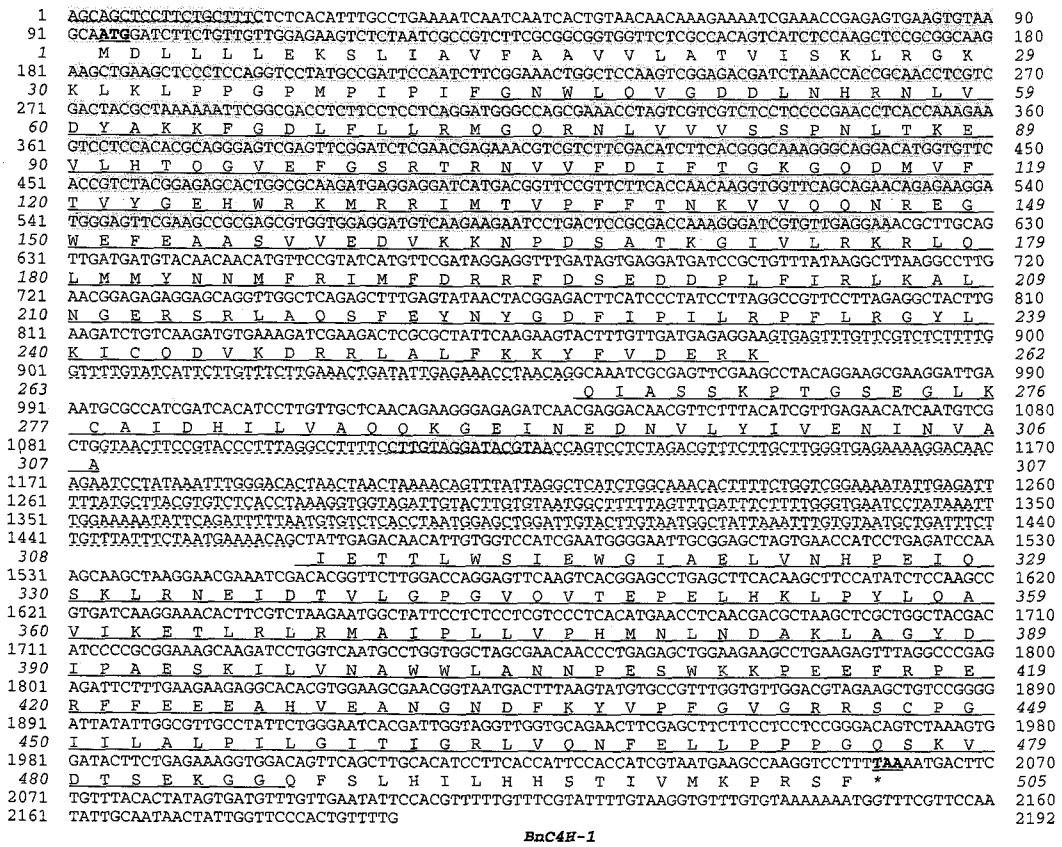

Fig. 1. Nucleotide sequences and deduced amino acid sequences of *BnC4H-1* and *BnC4H-2*. The start codon ATG and the stop codon TAA/TGA are in bold face and solid-underlined, while the introns are dash-underlined. The CpG island in *BnC4H-1* is in gray background, and the 19-bp conserved region in the 5' UTR, the 16-bp conserved region in the 2nd intron and the polyadenylation signal in the 3' UTR are in gray background and wave-underlined. The amino acid residues corresponding to the predicted conserved pfam00067 (p450) domain between F_{43} and G_{486} are solid-underlined.

Molecular characterization of nucleotide sequences of *BnC4H-1* and *BnC4H-2*

Basic parameters of *BnC4H-1* and *BnC4H-2*. The genomic sequence and full-length cDNA of *BnC4H-1* are 2192 bp and 1742 bp respectively (Fig. 1). When they were pairwise aligned, 2 introns (879-949 bp and 1084-1462 bp) were detected in this gene. They have standard GT……AG splicing sites with positions identical to those of *C4Hs* from *A. thaliana* and other plants. The full-length cDNA of *BnC4H-1* has a 93-bp leader sequence (5' UTR) and a 131-bp 3' UTR, between which is a 1518-bp ORF (including stop codon TAA). The G + C content of the ORF is 50.59%, while the non-coding regions have typically low G + C contents, e.g. 39.78, 31.30, 33.80 and 35.09% for 5' UTR, 3' UTR, intron 1 and intron 2 respectively.

The genomic sequence and full-length cDNA of *BnC4H-2* are 2108 bp and 1716 bp respectively (Fig. 1). The 2 introns (881-945 bp and 1080-1406 bp) also have standard GT…… AG splicing sites with positions corresponding to those of *BnC4H-1*. The 5' UTR, ORF and 3' UTR of *BnC4H-2* are 95 bp, 1518 bp (including stop codon TGA) and 103 bp, respectively. The G + C contents of ORF, 5' UTR, 3' UTR, intron 1 and intron 2 of *BnC4H-2* are 49.28, 40.00, 26.21, 32.31 and 35.47% respectively.

Homologies and origin of *BnC4H-1* and *BnC4H-2*. NCBI blastn indicated that the coding regions of *BnC4H-1* and *BnC4H-2* show high identities to known *C4H* tags from *Brassica* species and to *AtC4H* (**U71080** for gene and **NM_128601** for mRNA). They also show moderate identities to many non-cruciferous *C4H/CYP73A* genes, such as those from *Agastache rugosa* (**AY616436**), *Verbena* x *hybrida* (**AB234902**), *Parthenocissus henryana* (**DQ211885**), *Sorghum bicolor* (**AY034143**) and *Pinus taeda* (**AY764925**) etc. Among the known *Brassica C4H* tags, the 466-bp *C4H-BO-1* from *B. oleracea* (**AF230674**) shows 99% identities to *BnC4H-1* (only 1 bp of difference) in the 466-bp aligned region, while all other fragments including those from *B. napus* have local identities of less than 94%. These suggest that *BnC4H-1* is a novel *B. napus C4H* gene, which has no tag in the Genbank database and is undoubtedly transmitted from the parental species *B. oleracea*. On the other hand, the 314-bp *C4H-BN-7* from *B. napus* (**AF230673**) shows 98% identities to *BnC4H-2*. *C4H-BN-7* should be the exact tag of

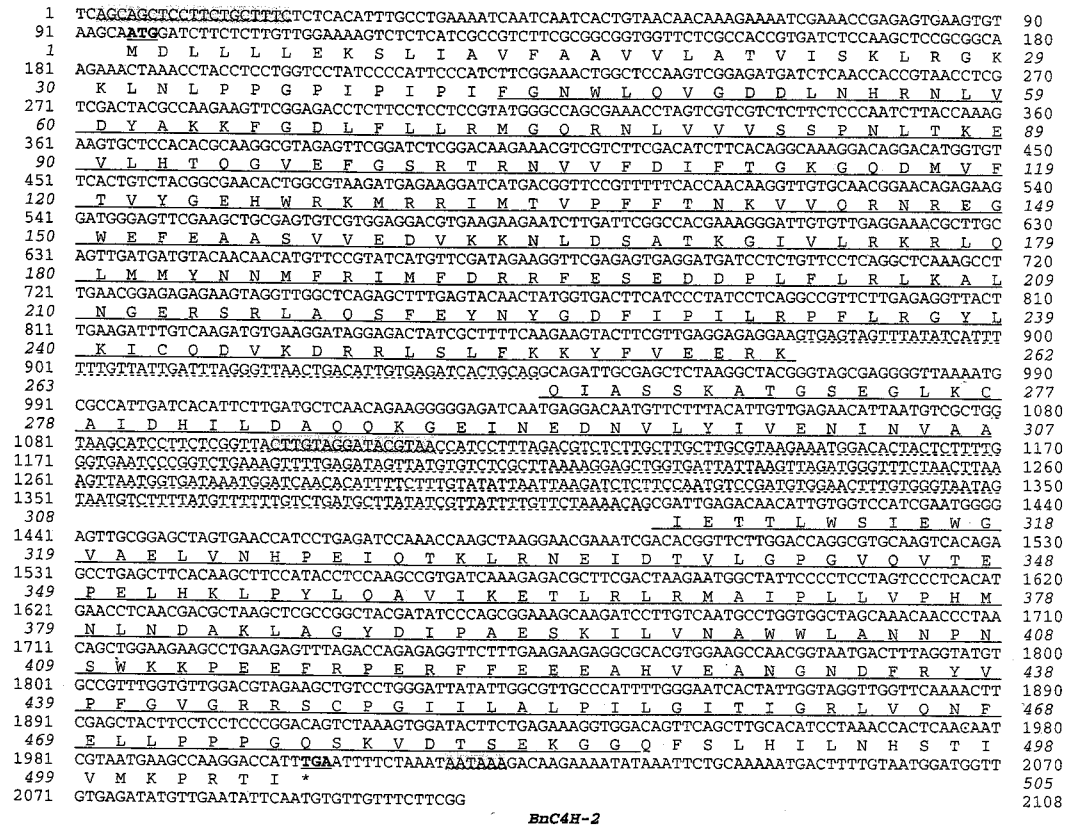

Fig. 1. Continued.

BnC4H-2 since *C4H-BN-7* was sequenced with "N" at the differed 6 bases. Though the 474-bp *C4H-BR-3* from *B. rapa* (**AF230680**) has 97% identities to *BnC4H-2*, it may not be the source gene of *BnC4H-2*, since non-neglectable divergences exist between them.

When pairwise-aligned on Vector NTI advance 9.0, *BnC4H-1* shows 80.6% and 87.4% identities to *BnC4H-2* on genomic and cDNA levels respectively. Their 5' UTRs are completely identical to each and their coding regions are of high identities (90.6%), while their introns and 3' UTRs are of low identities (59.7%, 53.8% and 47.5% for intron 1, intron 2 and 3' UTR respectively). *BnC4H-1* shows 75.9% and 83.1% identities to *AtC4H* on genomic and cDNA levels respectively. The identities of their ORF, 5' UTR, 3' UTR, intron 1 and intron 2 are 87.1, 59.1, 58.4, 63.5 and 45.1% respectively. *BnC4H-2* shows 77.3% and 83.1% identities to *AtC4H* on genomic and cDNA levels, respectively. The identities of their ORF, 5' UTR, 3' UTR, intron 1 and intron 2 are 87.0, 60.0, 53.3, 56.5 and 50.8% respectively.

Possible cis-elements of *BnC4H-1* and *BnC4H-2*. The 3' UTR of *BnC4H-2* contains a canonical polyadenylation signal A_{2017}ATAAA$_{2022}$, but none was detected in *BnC4H-1*. According to the new definition of CpG island (Takai and Jones, 2002), a 532-bp CpG island was predicted in *BnC4H-1* at the position A_{83}-A_{620} with a C + G-content of 55.1% and an Obs./Exp of 1.066. But due to lower G + C content, no CpG island was found in *BnC4H-2*. When the UTRs of *BnC4H-1* and *BnC4H-2* were aligned with those of *AtC4H*, several highly conserved regions were detected. The 5' UTRs of *BnC4H-1* and *BnC4H-2* are completely identical to each other in the corresponding 93-bp region, but even if in the coding regions there is no region as conserved as this region. On the other hand, though these two 5' UTRs are of low similarities to the 5' UTR of *AtC4H*, a 19-bp highly conserved region AGCAG CTCCTTCTGCTTTC was identified at the beginning of the 5' UTR of all the 3 genes. In this region, the two *B. napus* genes only show 1-bp difference to *AtC4H* (Fig. 2). Though most regions of the 2 introns of the 3 genes are of little conservation, but a highly conserved region was still detected at the beginning of the 2nd intron corresponding to T_{1109}-G_{1153} of *BnC4H-1* and T_{1097}-G_{1140} of *BnC4H-2*. Especially, within this region a 16-bp sequence CTTGTAGGATACGTAA, corresponding to 1112-1127 bp of *BnC4H-1* and 1100-1115 bp of *BnC4H-2*, is completely identical in the 3 genes (Fig. 2).

Conservation and structural features of the deduced BnC4H-1 and BnC4H-2 proteins

Basic properties of BnC4H-1 and BnC4H-2. The ORFs of *BnC4H-1* and *BnC4H-2* both encode a polypeptide of 505 amino acid residues. BnC4H-1 possesses a calculated molecular weight of 57.73 kDa and an isoelectric point (pI) value of 9.11, while BnC4H-2 is 57.75 kDa with a pI value of 9.13. L is the most abundant amino acid (11.09% and 11.49% for

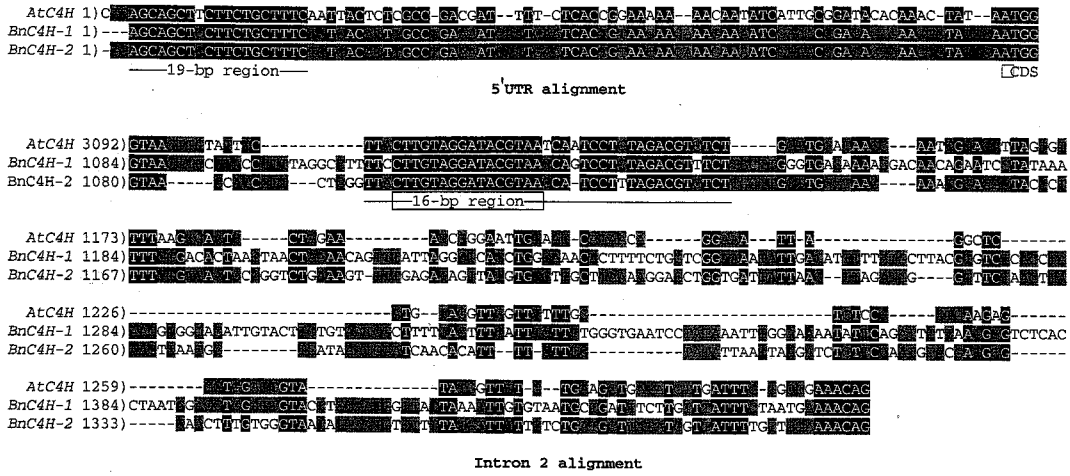

Fig. 2. Alignments of the 5' UTRs and the 2nd introns of *BnC4H-1*, *BnC4H-2* and *AtC4H*. Identical bases are in reverse display, with block similar bases in gray background and non-similar bases in no background. Conserved regions in the 5' UTR and the 2nd intron are marked. CDS indicates coding sequence after the start codon ATG.

BnC4H-1 and BnC4H-2 respectively), followed by V, K, E, I, R and G etc. Their basic amino acid contents (both are 13.27%) are higher than their acidic amino acid contents (both are 11.68%).

Homological analysis of BnC4H-1 and BnC4H-2. BnC4H-1 and BnC4H-2 show as high as 96.6% identities and 98.4% positives to each other on protein level. Among the 17 residues differed between them, 9 are substitutions by similar residues. SUPERFAMILY alignment (Madera *et al.*, 2004) revealed that BnC4H-1 and BnC4H-2 both belong to the cytochrome P450 family. NCBI blastp indicated that BnC4H-1 and BnC4H-2 show very wide similarities to C4Hs from other plants. When pairwise-aligned on whole molecule scale, BnC4H-1 shows identities/positives of 95.8%/98.0% to intra-family AtC4H (**AAB58355**), 85.9%/93.9% to dicot C4H from *Malus* x *domestica* (**AAY87450**), 58.4%/68.4% to another dicot C4H from *Mesembryanthemum crystallinum* (**AAD11427**), 75.4%/85.3% to monocot C4H from *S. bicolor* (**AAK54447**), and 77.7%/88.1% to gymnosperm C4H from *Ginkgo biloba* (**CAA70596**). The same trend is for BnC4H-2, which shows identities/positives of 95.4%/97.8% to **AAB58355**, 85.5%/93.1% to **AAY87450**, 58.6%/68.8% to **AAD11427**, 75.0%/84.8% to **AAK54447**, and 78.1%/87.9% to **CAA70596** respectively.

NCBI blastp also indicated that the two BnC4Hs show lower similarities to non-C4H P450s, such as flavonoid 3'-hydroxylase (F3'H), ferulate-5-hydroxylase (F5H), flavonoid 3'5'-hydroxylase (F3'5'H), *p*-coumaroyl CoA shikimate/quinate 3'-hydroxylase (C3'H) and (S)-N-methylcoclaurine 3'-hydroxylase ((S)-N-M3'H) etc. For example, the identities/positives of BnC4H-1 to *Antirrhinum majus* F3'H (**ABB53383**) are 29.1%/43.3%, to *Callistephus chinensis* F3'5'H (**AAG49299**) are 29.8%/43.8%, to *Coptis japonica* (S)-N-M3'H (**BAB12433**) are 29.1%/45.8%, to *Broussonetia papyrifera* F5H (**AAW50818**) are 27.3%/40.7%, and to *A. thaliana* C3'H (**NP_850337**) are 31.2%/48.7%, respectively. The identities/positives of BnC4H-2 to **ABB53383**, **AAG49299**, **BAB12433**, **AAW50818** and **NP_850337** are 29.5%/43.3%, 29.0%/43.3%, 29.4%/45.7%, 28.1%/41.6% and 31.2%/48.3%, respectively.

On the phylogenetic tree, BnC4H-1 and BnC4H-2 tightly sub-group with AtC4H, then with other C4Hs to form a highly homologous large group. While all other non-C4H P450s form another large group, though they also have certain similarities to BnC4H-1 and BnC4H-2 (Fig. 3).

Conserved domains/motifs and active site residues in BnC4H-1 and BnC4H-2. NCBI Conserved Domain (CD) search (Marchler-Bauer and Bryant, 2004) detected two conserved domains dominating most part of BnC4H-1 and BnC4H-2: pfam00067 (p450) and COG2124 (CypX). They both are conserved domains of cytochrome P450 proteins and have nearly overlapping locations in BnC4H-1: pfam00067 resides between F_{43} and G_{486} with a 95.4% alignment of the 461-residue CD-Length and a score of 296 bits, and COG2124 resides between Q_{48} and G_{475} with a 90.0% alignment of the 411-residue CD-Length with a score of 74.4 bits. In BnC4H-2, pfam00067 lies between F_{43} and V_{499} with a 98.0% alignment and a score of 299 bits, and COG2124 lies between Q_{48} and G_{475} with a 90.0% alignment and a score of 73.2 bits (Fig. 1). BnC4H-1 and BnC4H-2 have all the P450-featured motifs (Fig. 3), such as the haem-iron binding domain P_{439}FGVGRRSCPG$_{449}$, the T-containing binding pocket motif A_{306}AIETT$_{311}$, the E_{363}-R_{366}-R_{420} triad, and the hinge motif P_{34}PGP(M/I)PIP$_{41}$ necessary for optimal orientation of the enzyme (Chapple, 1998; Werck-Reichhart *et al.*, 2002).

Residues for enzymatic active sites of C4H may involve I_{109}, K_{113}, V_{118}, F_{220}, E_{301}, N_{302}, I_{303}, V_{305}, A_{306}, T_{310}, R_{366}, R_{368}, A_{370}, I_{371}, P_{372}, L_{374}, V_{375}, P_{376}, H_{377}, K_{484}, F_{488}, and L_{490} etc. They distribute in five signature motifs, i.e. substrate recognition sites (SRS), of C4H/CYP73A5: SRS1 (S_{100}RTRN VVFDIFTGKGQDMVFTVY$_{122}$), SRS2 (L_{216}AQSFEYNY$_{224}$),

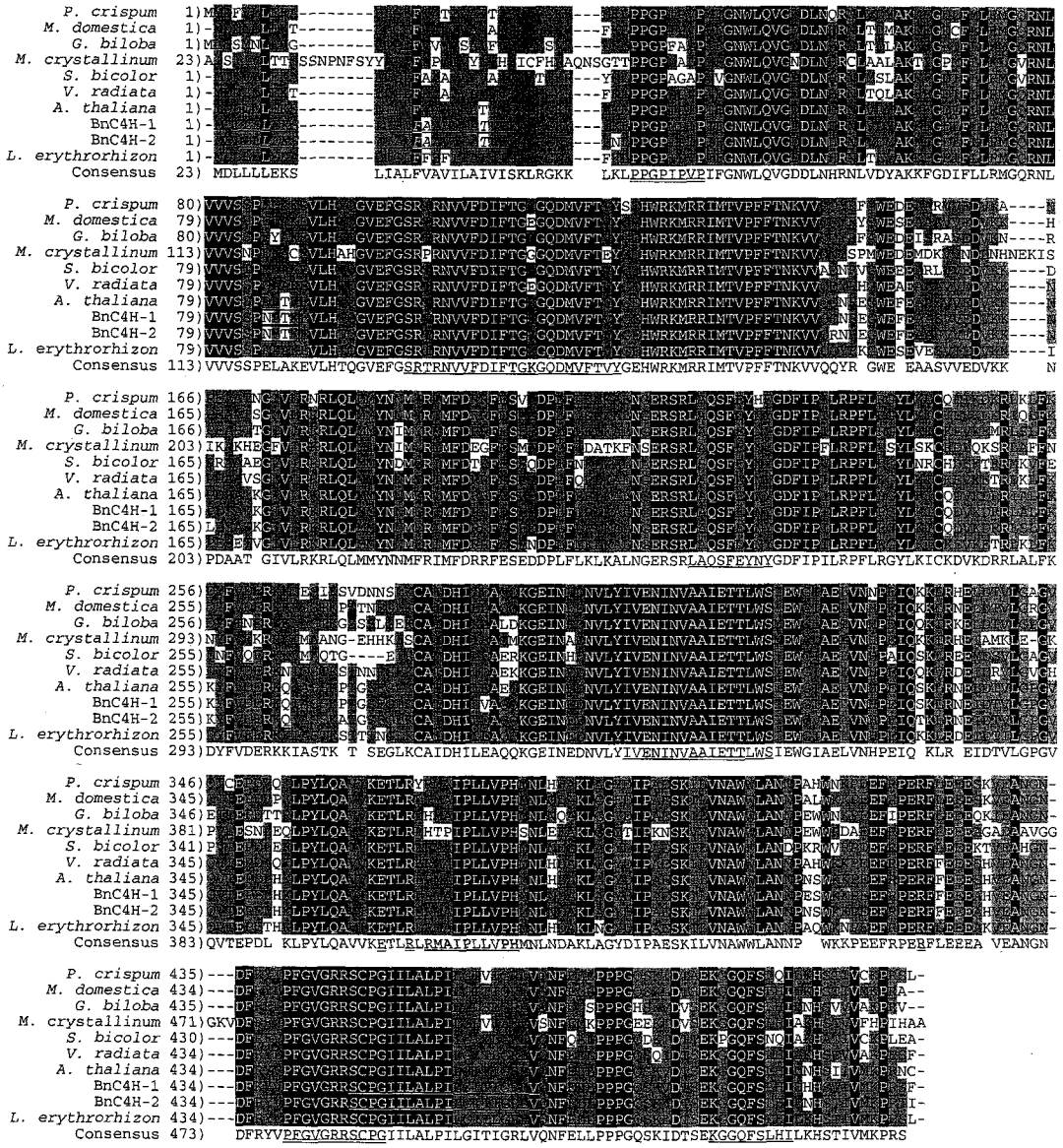

Fig. 3. Upper panel: multi-alignment of BnC4H-1 and BnC4H-2 with C4Hs from other plants. Besides BnC4H-1 and BnC4H-2, other C4Hs (Genbank accession numbers in parentheses) are from *Arabidopsis thaliana* (AAB58355), *Lithospermum erythrorhizon* (BAB71716), *Malus* x *domestica* (AAY87450), *Petroselinum crispum* (Q43033), *Vigna radiata* (AAA33755), *Ginkgo biloba* (AAW70021), *Sorghum bicolor* (AAK54447) and *Mesembryanthemum crystallinum* (AAD11427). In BnC4H-1 and BnC4H-2, the predicted signal peptide/anchor is in italics, while the N-terminal and the C-terminal transmembrane helices are underlined. In the consensus, double-underlined regions/residues indicate P450-featured motifs such as the hinge region, the T-containing binding pocket motif, the E_{363}-R_{366}-R_{420} triad and the haem domain, while single-underlines represent the 5 SRS regions in which the possible active site residues are in gray back-ground. Lower panel: phylogenetic tree of BnC4H-1 and BnC4H-2. C4Hs are the same as in the upper panel, while non-C4H P450s are: F3'Hs from *Antirrhinum majus* (ABB53383) and *Vitis vinifera* (BAE47004), F3'5'Hs from *Callistephus chinensis* (AAG49299) and *Gossypium hirsutum* (AAP31058), (S)-N-M3'Hs from *Coptis japonica* (BAB12433) and *Eschscholzia californica* (AAC39452), F5Hs from *Broussonetia papyrifera* (AAW50818) and *Medicago sativa* (ABB02161), and C3'Hs from *A. thaliana* (NP_850337) and *Sesamum indicum* (AAL47545), respectively. Calculations were done by Neighbor Joining method in AlignX of Vector NTI Advance 9.0. Calculated distance values are shown in parentheses following molecule names.

SRS4 (I_{299}VENINVAAIETTLWS$_{314}$), SRS5 (R_{368}MAIPLLVP H_{377}) and SRS6 (K_{484}GGQFSLHI$_{492}$), respectively (Hasemann *et al.*, 1995; Rupasinghe *et al.*, 2003; Schoch *et al.*, 2003). The SRS5 and SRS6 regions and the C-terminal end of the SRS4 region are important in contacting the aromatic rings of the substrates, and the SRS1 and SRS2 regions and the N-terminal end of the SRS4 are important in contacting the aliphatic regions of the substrates. Both BnC4H-1 and BnC4H-2 have all the five signature motifs with 100% identities to typical residues (Fig. 3). These results indicated

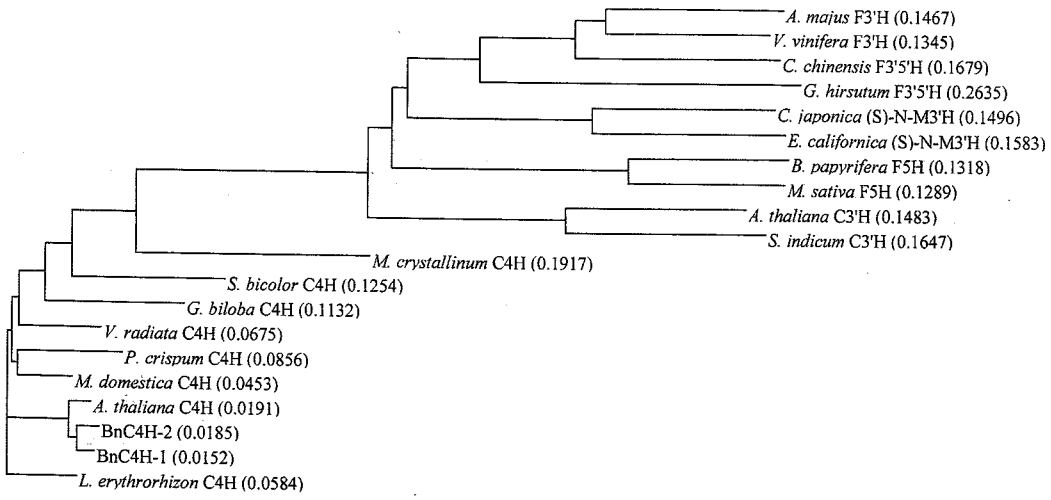

Fig. 3. Continued.

that BnC4H-1 and BnC4H-2 are no other than orthologous proteins of AtC4H (CYP73A5) and are most probably catalytically functional.

Possible post-translational modifications of BnC4H-1 and BnC4H-2. NetPhos 2.0 predicted 19 significant potential phosphorylation sites in BnC4H-1 (12 for S, 4 for T and 3 for Y) and 21 in BnC4H-2 (13 for S, 5 for T and 3 for Y), suggesting that phosphorylation may be a prerequisite for normal functioning of BnC4H-1 and BnC4H-2. NetNGlyc 1.0 (Blom *et al.*, 2004) and PROSITE predicted a potential agreement of 0.736 and 0.7362 for BnC4H-1 and BnC4H-2, respectively, to have an N-glycosylation site at position 85 (NLTK). Whether this site is really glycosylated *in vivo* needs experimental clues.

Signal peptide/anchor and subcellular localization of BnC4H-1 and BnC4H-2. SignalP 3.0 (Bendtsen *et al.*, 2004) predicted that BnC4H-1 has a probability of 0.408 to have a signal peptide and the probability for a signal anchor is 0.584, whereas 0.438 and 0.553 for BnC4H-2 (Fig. 3). Predotar (Small *et al.*, 2004) predicted endoplasmic reticulum (ER) scores of 0.83 and 0.86 for BnC4H-1 and BnC4H-2 respectively. Softberry-ProtComp 6.0 (**http://www.softberry. com/berry.phtml**) also definitely predicted them to be ER-membrane bound with scores of 3.1 and 3.0 respectively. WoLFPSORT suggested ER localization of them similar to At2g30490 (AtC4H). Based on above predictions and their sequence homologies with other C4Hs, these 2 C4H proteins are probably located in the endoplasmic reticulum, as already established for other C4Hs. But other location can not be excluded, since softwares like TargetP 1.1 predicted different results.

Both TMpred (Hofmann and Stoffel, 1993) and SOSUI (Mitaku *et al.*, 2002) predicted 2 strong transmembrane helices at both terminal regions of BnC4H-1 and BnC4H-2, and the positions and sequences are completely identical between the 2 proteins. In the TMpred results, the 22-residue N-terminal one is from L_3 to S_{24}, o-i oriented, with a score of 1666. The 20-residue C-terminal one is from S_{446} to V_{465}, i-o oriented, with a score of 1434 (Fig. 3). In the SOSUI results, the N-terminal 23-residue one is from L_3 to K_{25}, while the C-terminal 23-residue one is from S_{446} to F_{468}. The N-terminal transmembrane helix almost overlaps with the predicted N-terminal signal peptide/anchor (Fig. 3).

Secondary and tertiary structures of BnC4H-1 and BnC4H-2. Predicted by SOPMA (Geourjon and Deléage, 1995), the secondary structures of BnC4H-1 and BnC4H-2 are mainly composed of alpha helices (48.71% for both) and random coils (33.47% and 34.65%), while extended strands (12.67% for both) and beta turns (5.15% and 3.96%) also contribute. Alpha helices mainly distribute at the middle region and the N-terminus. In both proteins, the region between H_{125} and R_{366} is dominated by alpha helices (10 in BnC4H-1 and 11 in BnC4H-2) connected by random coils. Within this region, there is a 62-residue huge alpha helix from L_{202} to Q_{263} in BnC4H-1. In BnC4H-2 this huge helix is cut into two major helices by some random coils, but other 2 helices (H_{125}-F_{138} and K_{141}-K_{163}) found in BnC4H-1 have merged into a large helix (H_{125}-K_{163}) in BnC4H-2. The N-terminal large helix covers the predicted signal peptide/anchor and the N-terminal transmembrane helix. Extended strands mainly disperse at two regions: one is the ~100-residue C-terminal region, and another is the ~130-residue region between the N-terminal helix and the central helices. In these 2 regions, extended strands distribute in an interlaced manner with random coils and small alpha helices (Fig. 4).

The tertiary structures of BnC4H-1 and BnC4H-2 (Fig. 5) were predicted by SWISS-MODEL (Guex and Peitsch, 1997). The tertiary structures of BnC4H-1 and BnC4H-2 are very similar to the reported P450 crystal structure (Rupasinghe *et al.*, 2003), which is a globular protein. The haem is located in the center of the globular protein and is surrounded by several large alpha helices. The C4H-signature motifs SRS1, SRS2

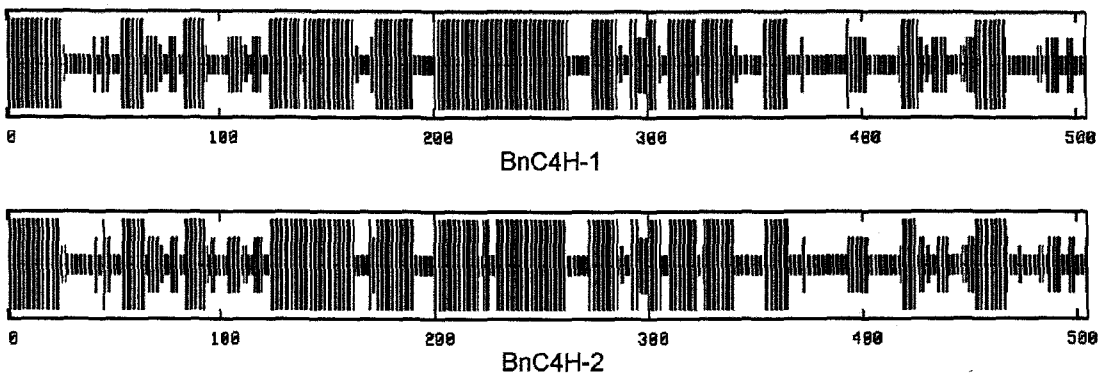

Fig. 4. Distribution of predicted secondary structures of BnC4H-1 and BnC4H-2. Four kinds of line bars in descending order in length represent alpha helix, extended strand, beta turn and random coil respectively. The numerals are residue counts along the whole proteins.

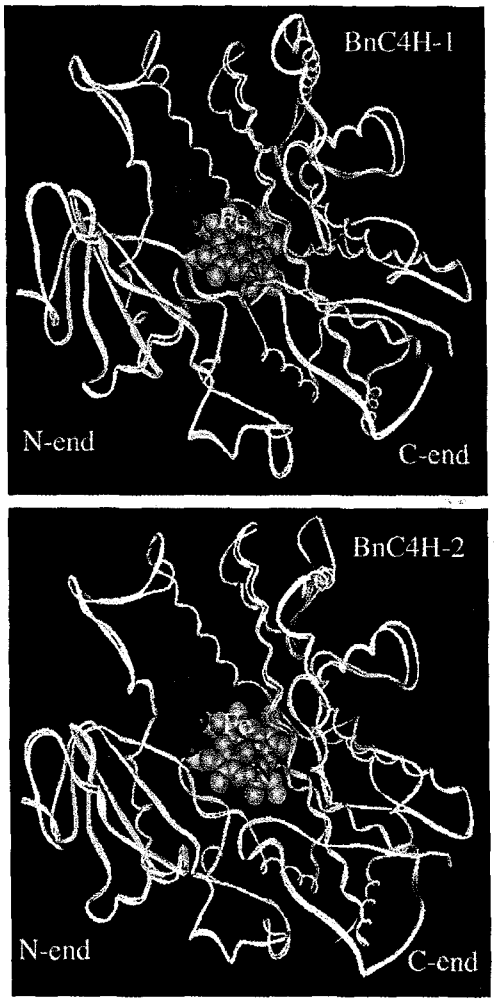

Fig. 5. Predicted tertiary structures of BnC4H-1 and BnC4H-2. O, N and Fe denote oxygen, nitrogen and iron atoms in the catalytic center respectively.

and SRS4 distribute in the helices 4, 8 and 10 (counting from N-terminus), respectively. The haem-iron binding motif P_{439} FGVGRRSCPG$_{449}$ locates in the SRS6. SRS motifs play important roles in constituting the enzymatic active site. As all other secondary structures predicted can be found, the N-terminal large helix is absent in the tertiary model.

Southern blot detection of *C4H* homologues in the genome of *B. napus*. After stringent hybridization and washing of Southern blot with *B. napus* genomic DNA, immunological detection showed that *Dra*I, *Eco*RI, *Eco*R*V* and *Hin*dIII digestions resulted in 5, 6, 8 and 9 hybridization bands respectively (Fig. 6). A few bands are quite weak, but they can still be identified as specific hybridization bands. Because all the 4 enzymes have no cutting site in the probe region even in the whole gene region of *BnC4H-1* and *BnC4H-2*, it is suggested that the *B. napus* genome may contain about 9 or more *C4H* members and some members have the same digestion maps for *Dra*I and *Eco*RV, respectively.

Transcription levels of *BnC4H-1* and *BnC4H-2* in various organs of *B. napus*. We adopted RT-PCR to detect isoform-specific expression of *BnC4H-1* and *BnC4H-2*. The results indicated that *BnC4H-1* and *BnC4H-2* have similar expression patterns in view of organ specificity, but differences are still obvious. The transcription of *BnC4H-1* can be distinctly detected in all analyzed organs except in 30 DAF seed. Its expression in the hypocotyl, stem, cotyledon, leaf, bud, flower and silique pericarp shows no great difference, but the expression in the root and seed is distinctly lower (Fig. 7). Expression of *BnC4H-2* can be detected in all the 11 organs analyzed including the 30 DAF seed. Its expression in the hypocotyl, stem, root, cotyledon, bud and flower is obviously higher than in the leaf, seed of all stages and silique pericarp, with the lowest still in the 30 DAF seed (Fig. 7). In order to eliminate experimental error, we repeated the RT-PCR three times and got very similar results.

Discussion

How many *C4H* genes in *Brassica* and Brassicaceae? The karyotype of *A. thaliana* has experienced a shrinking process featured by reciprocal translocations, inversions, chromosome

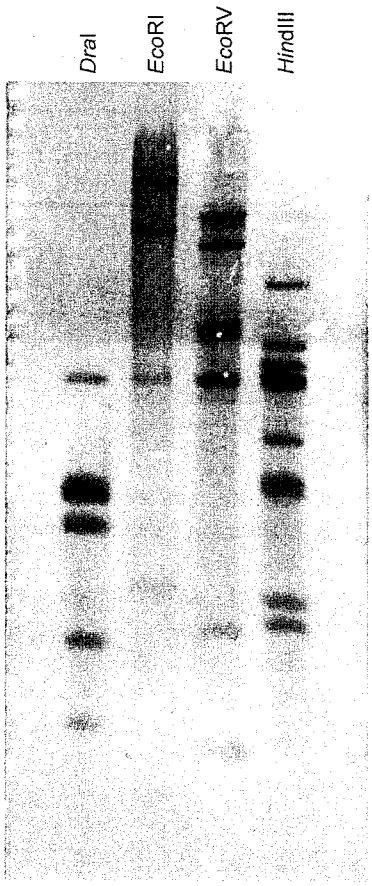

Fig. 6. Southern blot detection of homologous *C4H* members in *B. napus*.

fusions, and fragment lost (Lysak *et al.*, 2006). However, the genus *Brassica* has highly replicated genomes. Basic 'diploid' *Brassica* species are likely derived from hexaploid ancestry, and both recent and ancient polyploidization events generate a large number of genome rearrangements and novel genetic variation for important traits (Lukens *et al.*, 2004). Regions of the *A. thaliana* genome are triply present within the 'diploid' *Brassica* genomes (Cavell *et al.*, 1998; Lysak *et al.*, 2005). The genome of *B. napus*, an amphidiploid of *B. rapa* and *B. oleracea*, is more than 6 times of the *A. thaliana* genome (Schmidt *et al.*, 2001). Hence it is expected that in *B. napus* there might exist about 6 *C4H* genes, each orthologous to *AtC4H*.

In our study, two cDNAs and corresponding genomic sequences encoding C4H were isolated from *B. napus*. In a consensus genetic marker (ACGM) analysis, a pair of *AtC4H*-based conserved primers amplified 7, 4 and 3 *C4H* fragments from *B. napus*, *B. oleracea* and *B. rapa* respectively, and these numbers were considered reliable to represent the total *C4H* genes in the respective genomes (Fourmann *et al.*, 2002). But our Southern blot result, somewhat out of prediction, indicated that there might be as many as 9 or more *C4H* genes in *B. napus*. The G + C content of the 656-bp probe is 52.27%, while the Southern hybridization was stringently performed at 42°C with stringent washing. This excluded cross hybridization with non-*C4H* genes. If one gene has endonuclease cutting site(s) within the probe region, simultaneous detection of its two or more fragments seemed difficult under stringent conditions. Even if this overestimation phenomenon exists, there is a strong factor leading to underestimation of copy numbers in Southern blotting, i.e. identical digestion maps for two or more extremely homologous genes especially in the amphidiploid *B. napus* whose two parental species are also quite near in evolution.

Two important facts favor the high copy number assumption. First, *BnC4H-1* is just one-base different from *B. oleracea C4H-BO-1* (**AF230674**) in the alignable 466-bp region, indicating that a *B. napus* gene is basically unchanged from its donor gene from a parent species. This fact was also proved by Fourmann *et al.* (2002), who unambiguously assigned 43 out of the 102 *B. napus* genes to its parental species by sequencing. But in their research none of the 7 *B. napus C4H* tags was assigned to an explicit parental locus, and *vise versa* for the 7 parental *C4H* tags. This strongly suggests that, at least a part of, the 7 *B. napus C4H* tags have no receptor-donor relationship with the 7 parental-species *C4H* tags. This is to say that *C4H* gene numbers in *B. napus* should be more than 7 and also more than 4 and 3 in *B. oleracea* and *B. rapa* respectively. At least *BnC4H-1* is the 8th *C4H* gene in *B. napus* succeeding the 7 tags. Second, *B. oleracea C4H-BO-4* (**AF230677**) forms an almost triangle relationships with *AtC4H* and other known *C4H* tags/genes. Its identities to *AtC4H* are 80.9%, whereas just 83.7%-86.3% to all other *Brassica C4H* tags/genes (including *B. oleracea C4H* tags). In the phylogenetic tree, *C4H-BO-4* does not group with any other *Brassica C4H* genes/tags (Fig. 8A). On protein level, surprisingly, C4H-BO-4 is more divergent from other

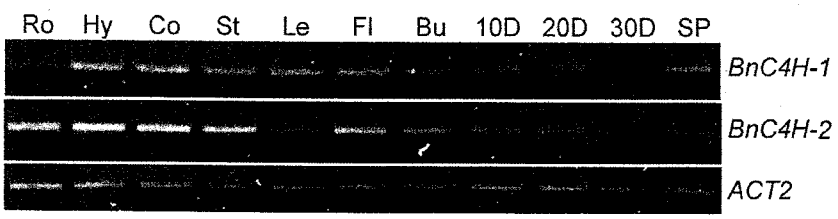

Fig. 7. RT-PCR detection of transcription levels of *BnC4H-1* and *BnC4H-2* in various organs of *B. napus*. Ro: root, Hy: hypocotyl, Co: cotyledon, St: stem, Le: leaf, Fl: flower, Bu: bud, 10D: seed of 10 d after flowering (DAF), 20D: seed of 20 DAF, 30D: seed of 30 DAF, and SP: silique pericarp.

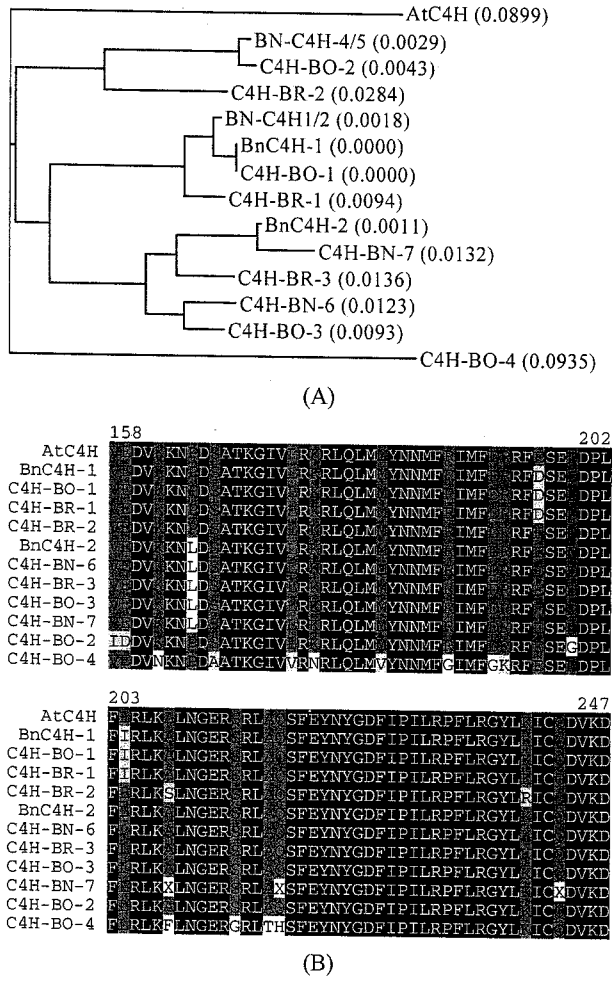

Fig. 8. *AtC4H* and *Brassica C4H* genes/tags: nucleotide phylogenetic tree (A) and amino acid alignment (B). Sequences and Accession numbers of the *Brassica C4H* tags can be found in the reference (Fourmann *et al.*, 2002) and on related website http://www.inra.fr/Internet/Produits/acgm/. For each gene/tag, a 278-bp fragment corresponding to the 278-bp *C4H-BR-2* is used for the phylogenetic analysis, while only the first 90 residues of the correspondingly deduced 92 amino acid residues are used in multi-alignment because many *Brassica C4H* tags contain "X" at the last two residues. The aligned protein region corresponds to V_{158}-D_{247} of AtC4H, BnC4H-1 and BnC4H-2. See Fig. 2 and Fig. 3 for related figure annotations.

Brassica C4Hs than AtC4H does (Fig. 8B). This indicates that the evolution of *C4H* genes in *Brassica* is more complicated than the triplication/hexaploidization assumption (Lysak *et al.* 2005). Perhaps after diverging from the ancestor of genus *Arabidopsis*, hexaploidization of the *Brassica* ancestor resulted in most of the known *Brassica C4H* genes/tags, but *C4H-BO-4* might be resulted from another duplication event prior to the triplication event. As to how many closely related homologues of *C4H-BO-4* exist in various *Brassica* or even *Brassiceae* species, further identification is necessary.

In the crucial common phenylpropanoid pathway, in sharp contrast to four *PAL* and four *4CL* genes, *A. thaliana* enigmatically contains only one *C4H* gene (Bell-Lelong *et al.*, 1997; Mizutani *et al.*, 1997). Pea and parsley also contain only one *C4H* gene (Frank *et al.*, 1996; Koopman *et al.*, 1999), but most plants contain a small family of *C4H* genes, e.g. 4 in rice (Nelson *et al.*, 2004), at least 2 in 'Valencia' orange (Betz *et al.*, 2001) and an undefined *C4H* family in mung bean (Mizutani *et al.*, 1993). Studies suggest that prior to the separation of monocots and dicots, or even earlier, the *C4H* gene has duplicated. Quite divergent classes, Class 1 and Class 2, of *C4H* genes have been identified in maize, French bean and 'Valencia' orange (Nedelkina *et al.*, 1999; Betz *et al.*, 2001). In Fig. 3, the separation of a dicot *M. crystallinum* C4H from all other dicot, monocot, even gymnosperm C4Hs, suggests the duplication of the *C4H* gene prior to the divergence of gymnosperm and angiosperm species. From this point of view, C4H in ancestral species of many dicot families including *Brassicaceae* may be encoded by more than one *C4H* genes. The evolution route of C4H may resemble those of PAL and 4CL. That is to say, C4H is encoded basically by multiple genes in higher plants, and the monogenic status of some plants is resulted from lost of duplicated genes. Exhaustive isolation of the whole *C4H* gene family in some *Brassicaceae* species will help to clarify whether the *Brassicaceae* ancestor was monogenic or multigenic at *C4H* and whether the monogenic feature of *AtC4H* is caused by gene deletion. It is also tempting to know whether the opposite evolution directions, "shrinking" in *Arabidopsis vs* "expanding" in *Brassica*, of the genome size especially of certain key functional genes (like *C4H*) have any correlation with their great differences in developmental traits such as biomass, plant height and seed size etc.

A few structural clues deserve further study. In this study, two possible *cis*-elements, one in the 5' UTR and another in the 2nd intron, were revealed. As calculations have indicated that, in sharp contrast with the coding regions, all the non-coding regions are basically of low conservation between *AtC4H* and the two *B. napus C4H* genes, but a 19-bp region AGCAGCTCCTTCTGCTTTC in the right beginning of the 5' UTR and a 16-bp sequence CTTGTAGGATACGTAA at the 5' of the 2nd intron are highly conserved (Fig. 2). The transcription initiation site of *AtC4H* has been located at C_1 in Fig. 2 (Bell-Lelong *et al.*, 1997), while the first bp of the two *B. napus C4H* genes are just 1-3 bp downstream of it. This strongly suggests that the transcription initiation sites of the two *B. napus C4H* genes conform to that of the *AtC4H* and conservation of certain proximal structures, e.g. right 5' region of 5' UTR, is a necessary determinative factor. This conserved region may participate in transcription regulation like a reported lepidopteran P450 gene (Petersen Brown *et al.*, 2004). Alternatively, the possible role of this 19-bp region may be involved in regulation of translation, but this possibility is obviously lower than the former one. The conservation of a 16-bp region at the 5' of the 2nd intron was also demonstrated

between *AtC4H* and the two *B. napus C4H* genes (Fig. 2). Possible role of this region may be involved in regulating transcription or transcript processing (Damert *et al.*, 1996), but experimental clues are required to give an affirmative answer. Furthermore, *BnC4H-1* and *BnC4H-2* differ from each other in CpG island and polyadenylation signal. Whether these differences have any relation to their different tissue specificities deserves clarification. The possible *cis*-elements revealed here provide useful motifs for elucidating the expression regulation of plant genes.

Another interesting fact is that the first ~100-bp regions of the two *B. napus C4H* genes are identical to each other (Fig. 2), while any other region including regions coding highly conserved C4H motifs does not show this high degree of conservation. Even if it is in demand for functional conservation, there is no need not to vary a base in the ~100-bp 5' UTR. This implies that a non-allelic fragment exchange/substitution might have occurred between these two genes. But since the tissue specificities of their transcription are different, the possible non-allelic fragment exchange/substitution did not impair the functional discrimination of their promoters.

BnC4H-1 and *BnC4H-2* encode typical C4H proteins, as they possess all the conserved motifs and active site residues required for C4H-type P450 proteins. All the varied residues in these two proteins locate at non-conserved sites at which C4H proteins from other species also show variations, but there still are a few residues in BnC4H-1 and BnC4H-2 showing less similarities to the typical residues. These residues are N_{85} in both proteins, R_{145} in BnC4H-2 and V_{284} in BnC4H-1. Whether these residues have any influence on the catalytic activity or substrate specificity of the two enzymes needs to be proved by functional identification.

Predictions on the signal peptide/anchor and subcellular localization of BnC4H-1 and BnC4H-2 by different softwares gave inconsistent results. C4H and other microsomal P450s have been extracted from microsomes of various organisms, and it has been presumed that microsomal P450s have an N-terminal hydrophobic helix which serves to anchor the enzyme to the ER membrane (Winkel-Shirley, 1999; Werck-Reichhart *et al.*, 2002). In BnC4H-1 and BnC4H-2 the N-terminal hydrophobic helix has been predicted with multiple identities, i.e. signal peptide with cutting site, signal anchor, and transmembrane helix, by different softwares. A second strong transmembrane helix was also predicted at the C-terminus. Another fact is that the predicted tertiary structures of the two proteins do not contain the first 30 residues, and no N-terminal helix can be seen in Fig. 5. Only a part of softwares predicted ER as the location site of these two proteins, while other softwares gave various results. Considering P450s in general, there is more than one report indicating effective or possible plastidic localization. These make it difficult to draw a definite conclusion on the properties of the N-terminal helix and subcellular localization of BnC4H-1 and BnC4H-2. But it is obvious that both the N-terminal and the C-terminal sequences of the two proteins do not show essential difference from those of AtC4H and most other C4Hs, so the location and topology of them should be similar to those of typical C4H proteins.

Conservation of protein structure and differentiation of tissue specificity of *C4H* genes. On whole nucleotide level the identities of *BnC4H-1* and *BnC4H-2* to *AtC4H* are only 75.9% and 77.3% respectively, but on ORF level they both show 83.1% identities to *AtC4H*. Nevertheless, on protein level BnC4H-1 and BnC4H-2 share surprisingly high identities (95.8%/95.4%) and positives (98.0%/97.8%) to AtC4H. As compared between *BnC4H-1* and *BnC4H-2*, though the identities are only 80.6% and 87.4% on genomic and cDNA levels respectively, the identities and positives on protein level are as high as 96.6% and 98.4% respectively. Through these data we can see that in Brassicaceae the *C4H* gene family is extremely conserved on protein level both orthologously and paralogously, and this conservation is combinatorially determined by coding region stability, codon degeneracy and similar amino acid substitution. Fig. 3 also indicates that at conserved motifs and active site residues, not only BnC4H-1 and BnC4H-2 are identical to each other, they also show little change as compared with AtC4H. Besides the primary structures, the secondary and tertiary structures of BnC4H-1 and BnC4H-2 are almost identical to each other.

On the other hand, expression patterns of *C4H* genes even within a species seem to be more differentiated. Though the wide-expression features of *BnC4H-1* and *BnC4H-2* resemble *AtC4H* (Bell-Lelong *et al.*, 1997), obvious differences also exist. Like *AtC4H*, *BnC4H-2* is also strongly expressed in root and stem, but the same strength was also found in hypocotyl, cotyledon and flower etc. *BnC4H-1* differs from *AtC4H* mainly in its low expression in root (Fig. 7). High-level expression of *BnC4H-1* and *BnC4H-2* in lowly lignified organs such as cotyledon, flower and bud suggests that they may play roles in non-lignification process, e.g. flavonoids biosynthesis. But since they are also highly expressed in hypocotyl and stem, their roles in lignification cannot be excluded.

Obvious complementation in tissue specificity can be found between the two isoforms. The transcript of *BnC4H-1* is not detectable in 30 DAF seed (old seed), while low-level expression of *BnC4H-2* is found in 30 DAF seed. *BnC4H-1* is obviously not the main isoform expressed in root, but the strong expression of *BnC4H-2* in root is a sufficient complementation. On the other hand, since *BnC4H-2* is not the main isoform in leaf and pericarp, high-level expression of *BnC4H-1* in these organs plays an important role. In other organs, such as hypocotyl, cotyledon, stem, flower, young- and middle-stage seed, the two genes have nearly the same levels of expression, perhaps reflecting a need for simultaneous strong functioning of them. Complementation and co-dominating both exist between these two isoforms as concerned with tissue specificity. This functional divergence in tissue specificity is obvious an evolution strategy to allocate the duplicated "redundant" family members especially in an

amphidiploid species like *B. napus*. In certain tissues, as assumed, isoforms of common phenylpropanoid pathway enzymes might be combined with certain branch pathway enzymes to form pathway-specific enzyme complex (Mizutani *et al.*, 1997; Winkel-Shirley, 1999). The features of tissue specificity of the two isoforms observed here favor this assumption. Maybe *BnC4H-1* and *BnC4H-2* have completed functional differentiation in certain tissues.

Other explanations for the differed tissue specificities include: 1) Mutations of certain tissue-specific cis-elements in the promoters make them low efficient in transcription in certain organs, and 2) In root, leaf, old seed and pericarp of *B. napus*, the full functioning of all the *C4H* isoforms is really somewhat redundant, so in these organs the advantageous isoform(s) were favored to be fully expressed and the disadvantageous one(s) were turned off gradually in the evolution. What is the actual fact needs to be experimentally identified.

Acknowledgments This research was supported by the Major Program of National Natural Science Foundation (30330400) and the Major Program of Chongqing Municipal Natural Science Foundation (8446) respectively.

References

Anterola, A. M. and Lewis, N. G. (2002) Trends in lignin modification: a comprehensive analysis of the effects of genetic manipulations/mutations on lignification and vascular integrity. *Phytochemistry* **61**, 221-294.

Barber, M. S. and Mitchell, H. J. (1997) Regulation of phenylpropanoid metabolism in relation to lignin biosynthesis in plants. *Int. Rev. Cytol.* **172**, 243-293.

Bell-Lelong, D. A., Cusumano, J. C., Meyer, K. and Chapple, C. (1997) Cinnamate-4-hydroxylase expression in *Arabidopsis* (Regulation in response to development and the environment). *Plant Physiol.* **113**, 729-738.

Benavente-Garcia, O., Castillo, J., Marin, F. R., Ortuno, A. and Del Rio, J. A. (1997) Uses and properties of *Citrus* flavonoids. *J. Agric. Food Chem.* **45**, 4505-4515.

Bendtsen, J. D., Nielsen, H., von Heijne, G. and Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340**, 783-795.

Betz, C., McCollum, T. G. and Mayer, R. T. (2001) Differential expression of two cinnamate 4-hydroxylase genes in 'Valencia' orange (*Citrus sinensis* Osbeck). *Plant Mol. Biol.* **46**, 741-748.

Blom, N., Sicheritz-Ponten, T., Gupta, R., Gammeltoft, S. and Brunak, S. (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* **4**, 1633-1649.

Cavell, A., Lydiate, D., Parkin, I., Dean, C. and Trick, M. (1998) Collinearity between a 30-centimorgan segment in *Arabidopsis thaliana* chromosome 4 and duplicated regions within the *Brassica napus* genome. *Genome* **41**, 62-69.

Chapple, C. (1998) Molecular-genetic analysis of plant cytochrome P450-dependent monooxygenases. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **49**, 311-343.

Damert, A., Leibiger, B. and Leibiger, I. B. (1996) Dual function of the intron of the rat insulin I gene in regulation of gene expression. *Diabetologia* **39**, 1165-1172.

Debeaujon, I., Nesi, N., Perez, P., Devic, M., Grandjean, O., Caboche, M. and Lepiniec, L., (2003) Proanthocyanidin-accumulating cells in *Arabidopsis* testa: regulation of differentiation and role in seed development. *Plant Cell* **15**, 2514-2531.

Di Carlo, G., Mascolo, N., Izzo, A. A. and Capasso, F. (1999) Flavonoids: old and new aspects of a class of natural therapeutic drugs. *Life Sci.* **65**, 337-353.

Dixon, R. A. and Paiva, N. L. (1995) Stress-induced phenylpropanoid metabolism. *Plant Cell* **7**, 1085-1097.

Dixon, R. A. and Steele, C. L. (1999) Flavonoids and isoflavonoids - a gold mine for metabolic engineering. *Trends Plant Sci.* **4**, 394-400.

Dixon, R. A., Lamb, C. J., Masoud, S., Sewalt, V. J. H. and Paiva, N. L. (1996) Metabolic engineering: prospects for crop improvement through the genetic manipulation of phenylpropanoid biosynthesis and defense responses - a review. *Gene* **179**, 61-71.

Fahrendorf, T. and Dixon, R, A. (1993) Molecular cloning of the elicitor-inducible cinnamic acid 4-hydroxylase cytochrome P450 from alfalfa. *Arch. Biochem. Biophys.* **305**, 509-515.

Fourmann, M., Barret, P., Froger, N., Baron, C., Charlot, F., Delourme, R. and Brunel, D. (2002) From *Arabidopsis thaliana* to *Brassica napus*: development of amplified consensus genetic markers (ACGM) for construction of a gene map. *Theor. Appl. Genet.* **105**, 1196-1206.

Frank, M. R., Deyneka, J. M. and Schuler, M. A. (1996) Cloning of wound-induced cytochrome P450 monooxygenases expressed in pea. *Plant Physiol.* **110**, 1035-1046.

Gabriac, B., Werck-Reichhart, D., Teutsch, H. and Durst, F. (1991) Purification and immunocharacterization of a plant cytochrome P450: the cinnamic acid 4-hydroxylase. *Arch. Biochem. Biophys.* **288**, 302-309.

Geourjon, C. and Deléage, G. (1995) SOPMA: Significant improvement in protein secondary structure prediction by consensus prediction from multiple alignments. *Cabios* **11**, 681-684.

Guex, N. and Peitsch, M. C. (1997) SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* **18**, 2714-2723.

Harakava, R. (2005) Genes encoding enzymes of the lignin biosynthesis pathway in *Eucalyptus*. *Genet. Mol. Biol.* **28**, 601-607.

Harborne, J. B. and Williams, C. A. (2000) Advances in flavonoid research since 1992. *Phytochemistry* **55**, 481-504.

Hasemann, C. A., Kurumbail, R. G., Boddupalli, S. S., Peterson, J. A. and Deisenhofer, J. (1995) Structure and function of cytochromes P450: a comparative analysis of three crystal structures. *Structure* **3**, 41-62.

Heneen, W. K. and Brismar, K. (2001) Maternal and embryonal control of seed colour by different *Brassica alboglabra* chromosomes. *Plant Breed.* **120**, 325-329.

Hofmann, K. and Stoffel, W. (1993) TMbase - A database of membrane spanning proteins segments. *Biol. Chem. Hoppe-Seyler* **374**, 166.

Jaakola, L., Pirttilä, A. M., Halonen, M. and Hohtola, A. (2001) Isolation of high quality RNA from bilberry (*Vaccinium myrtillus* L.) fruit. *Mol. Biotechnol.* **19**, 201-203.

Koopman, E., Logemann, E. and Hahlbrock, K. (1999) Regulation and functional expression of cinnamate 4-hydroxylase from parsley. *Plant Physiol.* **119**, 49-55.

Le Marchand, L. (2002) Cancer preventive effects of flavonoids: a review. *Biomed. Pharmacother.* **56**, 296-301.

Lukens, L., Quijada, P., Udall, J., Pires, J. C., Schranz, M. E. and Osborn, T. C. (2004) Genome redundancy and plasticity within ancient and recent *Brassica* crop species. *Biol. J. Linnean Soc.* **82**, 665-674.

Lysak, M. A., Berr, A., Pecinka, A., Schmidt, R., McBreen, K. and Schubert, I. (2006) Mechanisms of chromosome number reduction in *Arabidopsis thaliana* and related *Brassicaceae* species. *Proc. Natl. Acad. Sci. USA* **103**, 5224-5229.

Lysak, M. A., Koch, M. A., Pecinka, A. and Schubert, I. (2005) Chromosome triplication found across the tribe *Brassiceae*. *Genome Res.* **15**, 516-525.

Madera, M., Vogel, C., Kummerfeld, S. K., Chothia, C. and Gough, J. (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res.* **32**, 235-239.

Manthey, J. A., Guthrie, N. and Grohmann, K. (2001) Biological properties of citrus flavonoids pertaining to cancer and inflammation. *Curr. Med. Chem.* **8**, 135-153.

Marchler-Bauer, A. and Bryant, S. H. (2004) CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.* **32**, 327-331.

Mitaku, S., Hirokawa, T. and Tsuji, T. (2002) Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces. *Bioinformatics* **18**, 608-616.

Mizutani, M., Ohta, D. and Sato, R. (1997) Isolation of a cDNA and a genomic clone encoding cinnamate 4-hydroxylase from *Arabidopsis* and its expression manner in planta. *Plant Physiol.* **113**, 755-763.

Mizutani, M., Ward, E., DiMaio, J., Ohta, D., Ryals, J. and Sato, R. (1993) Molecular cloning and sequencing of a cDNA encoding mung bean cytochrome P450 (P450C4H) possessing cinnamate 4-hydroxylase activity. *Biochem. Biophys. Res. Comm.* **190**, 875-880.

Nedelkina, S., Jupe, S. C., Blee, K. A., Schalk, M., Werck-Reichhart, D. and Bolwell, G. P. (1999) Novel characteristics and regulation of a divergent cinnamate 4-hydroxylase (CYP73A15) from French bean: engineering expression in yeast. *Plant Mol. Biol.* **39**, 1079-1090.

Nelson, D. R., Schuler, M. A., Paquette, S. M., Werck-Reichhart, D. and Bak, S. (2004) Comparative genomics of rice and *Arabidopsis*. Analysis of 727 cytochrome P450 genes and pseudogenes from a monocot and a dicot. *Plant Physiol.* **135**, 756-772.

Petersen Brown, R., Berenbaum, M. R. and Schuler, M. A. (2004) Transcription of a lepidopteran cytochrome P450 promoter is modulated by multiple elements in its 5' UTR and repressed by 20-hydroxyecdysone. *Insect Mol. Biol.* **13**, 337-347.

Rechards, E. J. (1995) Preparation and analysis of DNA; in *Short Protocol in Molecular Biology*, Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A. and Struhl, K. (eds.), pp. 36-38, John Wiley and Sons, New York, USA.

Rupasinghe, S., Baudry, J. and Schuler, M. A. (2003) Common active site architecture and binding strategy of four phenylpropanoid P450s from *Arabidopsis thaliana* as revealed by molecular modeling. *Protein Eng.* **16**, 721-731.

Russell, D. W. (1971) The metabolism of aromatic compounds in higher plants. Properties of cinnamic acid 4-hydroxylase of pea seedlings and some aspects of its metabolic and developmental control. *J. Biol. Chem.* **246**, 3870-3778.

Sambrook, J. and Russell, D. W. (2001) Molecular Cloning: A Laboratory Manual, 3rd ed. Cold Spring Harbor Laboratory Press, New York, USA.

Schmidt, R., Acarkan, A., Boivin, K. (2001) Comparative structural genomics in the *Brassicaceae* family. *Plant Physiol. Biochem.* **39**, 253-262.

Schoch, G. A., Attias, R., Le Ret, M. and Werck-Reichhart, D. (2003) Key substrate recognition residues in the active site of a plant cytochrome P450, CYP73A1. Homology guided site-directed mutagenesis. *Eur. J. Biochem.* **270**, 3684-3695.

Seidman, C. E., Struhl, K. and Sheen, J. (1995) *Escherichia coli*, plasmids, and bacteriophages; in *Short Protocols in Molecular Biology*, Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A. and Struhl, K. (eds.), pp. 22-24, John Wiley and Sons, New York, USA.

Small, I., Peeters, N., Legeai, F. and Lurin, C. (2004) Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics* **4**, 1581-1590.

Takai, D. and Jones, P. A. (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. USA* **99**, 3740-3745.

Teutsch, H. G., Hasenfratz, M.-P., Lesot, A., Stoltz, C., Garnier, J.-M., Jeltsch, J.-M., Durst, F. and Werck-Reichhart, D. (1993) Isolation and sequence of a cDNA encoding the Jerusalem artichoke cinnamate-4-hydroxylase, a major plant cytochrome P450 involved in the general phenylpropanoid pathway. *Proc. Natl. Acad. Sci. USA* **90**, 4102-4107.

Weisshaar, B. and Jenkins, G. I. (1998) Phenylpropanoid biosynthesis and its regulation. *Curr. Opin. Plant Biol.* **1**, 251-257.

Werck-Reichhart, D., Bak, S. and Paquette, S. (2002) Cytochromes P450; in *The Arabidopsis Book*, Somerville, C. R. and Meyerowitz, E. M. (eds.), pp. 1-28, American Society of Plant Biologists, Rockville, USA.

Winkel-Shirley, B. (1999) Evidence for enzyme complexes in the phenylpropanoid and flavonoid pathways. *Physiol. Plant.* **107**, 142-149.