Genomics Grand for Diversified Plant Secondary Metabolites

Xin FANG[1,2], Chang-Qing YANG[2], Yu-Kun WEI[1], Qi-Xia MA[1],
Lei YANG[1,2], Xiao-Ya CHEN[1,2]*

(1 *Plant Science Research Center*, *Shanghai Chenshan Botanical Garden*, Shanghai 201602, China;
2 *Institute of Plant Physiology and Ecology*, *Shanghai Institutes for Biological Sciences*,
Chinese Academy of Sciences, Shanghai 200032, China)

Abstract: Plants can generate an overwhelming variety of structurally diversified organic compounds called secondary metabolites. These compounds usually perform interesting biological activities and important functions in influencing interactions between plants and other organisms. They are also widely utilized as pharmaceuticals, insecticides, dyes, flavors and fragrances. Plant genome sequencing, transcriptome and metabolome analyses have provided huge amounts of data to explain the great diversity of secondary metabolites. This knowledge in turn will help us better understand their ecological role and is a creating novel tool for genetic engineering of plant secondary metabolism.

Key words: Plant secondary metabolites; Biosynthetic pathways; Genomics; Chemical diversity; Bioinformatics

CLC number: Q 946.8　**Document Code**: A　**Article ID**: 2095-0845(2011)01-053-12

Plants produce structurally diversified secondary metabolites which are distributed differentially among limited taxonomic groups. Recent estimates are that a single plant species can produce 5 000 to 25 000 different compounds and that the over 100 000 known chemical structures represent a fraction of the total in the plant kingdom (Trethewey, 2004). They have been extensively utilized as dye, spices, glue, perfume and drugs for centuries by human beings. The early investigations of these organic compounds were mainly undertaken by organic chemists who developed separation techniques, spectroscopic approaches and synthetic methodologies to elucidate their structures, the complex biosynthetic pathways of most of these compounds, although frequently deduced, were unveiled during that time. There is no doubt that characterizations in chemistry have helped us to better utilize these secondary metabolites. However, for a long time, most plant biologists treated them as waste products of primary metabolism and paid little attention to them as chemical treasures as they do not appear to contribute to plant growth and development directly. Only recently, plant secondary metabolites were considered to influence ecological interactions between the plant and its environments in many aspects, such as inter- and intra-plant priming, direct and indirect defense, allelopathy, attraction, UV-B protection, responses to temperature, drought or nutrient stresses (Chen *et al.*, 2009). These ecological functions are of vital importance for plant survival and have attracted ever-increasing attention from plant biologists.

The identification and classification of the complex and diversified structures of plant secondary metabolites and the enzymes involved in the corresponding pathway are the basis for the development of organic chemistry and plant ecology. However, the natural products obtained by traditional chemistry methods and enzymes by biochemistry methods usually could not comprehensively account for the biosynthetic capacity of the organism.

Genome sequencing has opened a new era of life science research. Plant genomics provides both a

* Author for correspondence; E-mail: xychen@sibs.ac.cn

Received date: 2010-12-17, Accepted date: 2010-12-30

powerful tool and a rich source of data to dissect in order to better understand the origin of the great diversity of secondary metabolites. A plant genome may harbor more genes responsible for secondary metabolites than previously estimated. For example, of the 120 Mb Arabidopsis genome, 32 were annotated to code for terpene synthases, although functions of most of them remain to be experimentally characterized (Aubourg *et al.*, 2002). These genes were suggested to derive from individual gene tandem duplication and large-scale genome segmental duplication events. Plants could also create chemical diversity by using a matrix pathway, mix genes, and transcription factors without createing new genes. Genome mining could uncover compounds that have not been characterized by classical natural product isolation methods. We will then discuss the genomics grand for diversified plant secondary metabolites.

1 Major groups of plant secondary metabolites

Plant secondary metabolites arrive from limited precursor molecules, shared with plant primary metabolism, and are further transformed into complex and specific products through enzymatic catalyzed reactions in the plant. Based on biosynthetic origins, these diversified compounds can be divided into three major categories: terpenoids, alkaloids, and phenylpropanoids/allied phenolic compounds. There are other minor groups which cannot be included in these three groups, such as amines, cyanogenetic, glycosides, glucosinolates, acetlenes and psoralens.

1.1 Terpenoids

Terpenoids represent the most abundant and structurally diverse group of plant secondary metabolites, in which more than 36 000 individual structures have been identified. The diversified structures of terpenoids derive from a sequential assembly of five-carbon building blocks called isoprene units (C_5H_8), the so called biogenetic isoprene rule. Accordingly, one, two, three, four, five and six units construct hemiterpenes, monoterpenes, sesquiterpenes, diterpenes, sesterterpenes and triterpenes, respectively. The isoprene units are often joined in a “head-to-tail” and “head-to-head” fashions and a few are formed by head-to-middle fusions. After the formation of the basic terpenoid skeletons, subsequent modifications, including oxidation, reduction, isomerization, conjugation, and even degradation, lead to various structures of terpenoids. For example, steroids, a group of physiologically important natural products including cholesterol, ergosterol, and androgen, are degraded triterpenes.

The most typical hemiterpene is isoprene, a volatile product, which is released from photosynthetically active tissues. The monoterpenes and sesquiterpenes are components of volatiles and are widely used as flavors and perfumes. The diterpenes are mainly found in resin and triterpenes mainly occur as saponins. There are more than 30 basic monoterpene skeletons which could be divided into the acyclic (such as myrcene and citronellol), monocyclic (including the camphene, isocamphane and ionone types), bicyclic (including the thujane, carane, pinane, camphene, isocamphane, and fenchane types), and the iridoids (Grayson, 2000). The reported ecological functions of monoterpenes include attraction, allelopathy and defense, and some monoterpenes play different functions in different plants. The basic skeletons of sesquiterpenes are most abundant in the terpenoids, of which there were 60 known types in 1995 (Fraga, 1997). According to the structures, sesquiterpenes also could be grouped into the acyclic (such as farmesane), monocyclic (including the bisabolane, germacrane, humulane, and elemane types), bicyclic (including the caryophyllane, cadinane, eremophilane, valeriane, guaiane, eudesmane, illudane, acorane, himachanlane, laurane, chamigrane, and neopinguisene types), tricyclic (including the aristolane, maliane, thujopsane, proto-illudane, hirsutane, cedrane, marasmane, cubebane, longifolane, clovane, and junicedrane types). Similary, diterpenes also have acyclic and cyclic skeletons. The acylic diterpenes include such compounds as geranylgeranilo, the monocyclic diterpenes include cembrene, the bicyclic structures are represented by the labdane and clerodane types, the tricyclic by the pimarane and abietane types, and the tetracyclic by the kauranes, gibberellin, grayanane and phorbol types. Other important diterpenoids include taxanes and ginkgolides (Hanson, 2004). The ecological functions of sesquiterpenes, diterpenes and triterpenes are believed to participate in direct or indirect defense of the plant against herbivores and microbial pathogens. Triterpenoid are synthesized from the cyclization of squalene and most of them have tetracyclic and pentacyclic structures. The lanostane, dammarane, protostane, cucurbitane, sipholane, limonoid and quassinane types are the most common tetracyclic triterpenoids, whilst the oleanane, ursane, lupane, hopane, serratabe, malabaricane and ferane types are the most abundant pentacyclic triterpenoids (Connolly and Hill, 2010). Many terpenoids are of great medicinal vaule, for example, the sesquiterpene lacone artemisinin and the diterpene alkaloid taxol are drugs used for the treatments of malaria and cancer, respectively.

1.2 Alkaloids

Alkaloids are a group of alkaline, have a low molecular weight, and are nitrogen-containing compounds with at least one ring with the nitrogen atoms usually present in rings. They are the most widely distributed secondary metabolites and are found not only in plants, but also in microorganism animals, and play an important role in plant defense systems. Alkaloid-containing plants were, for mankind, the original "material medica" and many are still in use today as prescription drugs, such as vinblastine, quinine, atropine, and camptothecin. There are more than 12 000 alkaloids reported from 100 families of plants, and they are especially abundant in families of Leguminosae, Solanaceae, Manispermaceae, Papaveraceae, Ranurculaceae, and Berberidaceae (Buchanan *et al.*, 2000). Alkaloids seldom co-exist with terpenoids and volatile oils in the same plant and most of them exist in the form of salts in the plant, except for some with weak basis.

Aalkaloids can be classified on the basis of the plants from which they were isolated, the chemical structures, and the biosynthetic origins, and the last has an obvious advantage of reflecting the relationship between biosynthetic pathways and chemical structures. Alkaloids could thus be further classified into three groups according to their biosynthesis origin: true alkaloid, protoalkaloid, and pseudoalkaloid (Dewick, 2002). True alkaloids are formed from L-amino acids, such as tryptophan (Trp), tyrosine (Tyr), lysine (Lys), and arginine (Arg) (Dewick, 2002). For example, pyrrolizidine, pyrrolizidine and tropand alkaloids are formed from ornithine; piperidine, indolizidine and quiolizidine alkaloids (mainly found in Leguminosae) from lysine; quinoline, quinazoline and acridine alkaloids from anthranilic acid. Simple isoquinoline, benzylisoqunoline, bisbenzylisoquinoline, aporphine, berberine and protoberberine, protopine, emetine, α-naphtha-phenanthridine and morphine alkaloids are generated from Tyr; simple β-carboline, monoterpenoid indole and ergot alkaloids are originated from tryptophan. Protoalkaloids, such as ephedrine, pseudoephedrine and capsaicinoid, are a group of aromatic amine originating from phenylalanine (Phe) but with the nitrogen atom located inside the chain instead of the ring (Dewick, 2002). The precursors of pseudoalkaloids are not L-amino acids, and the nitrogen atoms are introduced into the structure in a later stage in the biosynthetic pathway, despite the location of nitrogen atom in ring (Dewick, 2002). These types of products include terpenoid alkaloids and steroidal alkaloids.

1.3 Phenylpropanoids

Phenylpropanoids/allied phenolic compounds contain at least one aromatic ring with one or more hydroxyl groups, and the majority of which are formed through shikimate/arogenate pathways alone or in combination with acetate/melonate pathways. More than 10 000 plant phenolic structures have been reported. Phenolics are of great importance as cellular support materials, such as lignins present in the cell wall, for mechanical support and as barriers against microbial invasion. The most magnificent function of the phenolic flavonoids, especially the anthocyanins, together with flavones and flavonols as co-pigments, is their contribution to flower and fruit colors (Dey and Harborne, 1997).

On the basis of their chemical structures and biosynthetic pathway, phenylpropanoids/allied phenolic compounds could be divided into following groups: lignins, lignans, coumarins, flavonoids, tannins, stibenes, styrylpyrones and arylpyrones. Lignins and lignans, two biosynthetically related secondary metabolites, are polymer, oligomer or dimmer of phenylpropanoids, respectively. Most lignans are optically active, but all isolated lignins are not. Cinnamyl alcohol, cinnamic acid, propenyl phenol and allyl phenol are the phenylpropanoid units consisting of lignans and their coupling modes are relatively few comparing to the several thousand known lignans. There are still no methods available for isolating lignins in their native state that do not markedly alter the original structure of the biopolymers during dissolution, and no methods of completely and effectively breaking down the biopolymers for structure analysis. The real structure of lignins remains a question, but we do know that gymnosperm lignins are primarily derived from coniferyl alcohol, and to a lesser extent, *p*-coumaryl alcohol, whereas angiosperms contain coniferyl and sinapyl alcohols in roughly equal proportions.

The structural feature of coumarins is a benzopyranone core, and they could be divided into simple coumarins, linear furanocoumarins, angular furanocoumarins, pyranocoumarins, and pyromesubstituted coumarins. There are more than 1 500 coumarins found in more than 800 species, occurring in the seed coats, fruits, flowers, roots, leaves and stems, although in general the greatest concentrations are found in fruits and flowers.

The structure of flavonoids contains a C_6-C_3-C_6 core which is two benzene (A ring and B ring) linked through a three carbons bridge or a pyran ring (C ring). They are biosynthesized through the condensation of three molecules of acetate-derived malonyl-CoA and one molecule of *p*-coumaryl-CoA. According to the oxidation of the C_3 chain and the linkage of the B ring, flavonoids could be classified into such subgroups as chalcones, aurones, flavonones, isoflavonoids, flavones, flavonols, leucoanthocyanidins, xanthones, aurones, furanochromones, homoisoflavones, phenylchromones, catechins, and anthocyanins. They could occur as monomers, dimmers and higher oligomers in most plant tissues, often in vacuoles. They are also found as mixtures of colored oligomeric/polymeric components in various heartwoods and barks. The coupling of flavoids could produce condensed tannins which add a distinct bitterness or astringency to the taste of certain plant tissues and function as antifeedants. In certain plant species cinnamoyl-CoA and malonyl-CoA pathways could also undergo condensation reactions to yield the corresponding stilbenes, styrylpyrones, and arylpyrones. The stilbene combretastatin has important antineoplastic activities, and resveratrol, present in red wine, helps suppress tumor formation.

2 Biosynthetic origins of secondary metabolism pathways

Secondary metabolites are derived from primary metabolism pathways. In many cases, primary and secondary metabolites cannot readily be distinguished on the basis of precursor molecules, chemical structures, or biosynthetic origins. As we have mentioned above, the investigation of the biosynthesis of secondary metabolism was mainly carried out by organic chemists in the early twenty century. The "isoprene rule" proposed by Otto Wallach, "biogenetic isoprene rule" by Leopold Ruzicka, and the alkaloid biosynthesis pathway suggested by Sir Robert Robinson, Clemens Schöpf, Ernst Winterstein and Georg Trier were a few of the most famous examples. From 1950's to 1970's, the establishment of precursor-feeding experiments and suspension cultures of plant cells allowed scientists to investigate the biosynthesis of secondary metabolites at the enzyme level, although this was typically limited to a small group of proteins that were abundant, stable, and soluble and for which substrates were commercially available. Since the 1980s, the ability to isolate mutant plants defective in multiple steps of biochemical pathways, along with the expression of plant proteins in microbes to assay enzyme activities *in vitro* or to complement existing mutations in the microorganism, led to hundreds of plant enzymes being identified and cloned. Once full-length cDNAs became available, the problem of obtaining enough protein for detailed enzymatic analysis could be circumvented by expression and purification of target proteins produced in microbes and even in animal cells. By using these methods, the biosynthesis pathways of some terpenoids, lignans, lignins and flavonoids have been characterized. As for alkaloids, at least eight pathways have been elucidated for the enzyme and gene level, including ajmaline, vindoline, berberine, corydaline, macarpine, morphine, berbamunine, and scopolamine (Ziegler and Facchini, 2008). The secondary metabolites are often stored and even formed in special structures of plants. The hydrophilic metabolites are often stored in vacuole, laticifers, and apoplast or cell walls, whilst the lipophilic compunds in cuticles, trichomes, resin ducts, laticifers, oil cells, and plastid membranes.

2.1 Biosynthesis of terpenoids

The investigation of the biosynthesis of terpenoids revealed that the real precursors of terpenoids are not isoprene, but isopentenyl diphosphate (IPP) and its allylic isomer dimethylallyl diphosphate (DMAPP), and the geranyl diphosphate (GPP), farmesyl diphosphate (FDP) and geranylgeranyl diphosphate (GGPP) produced by repetitive additions of IPP. The elaboration of these allylic prenyl diphosphates by specific terpenoid syntheses yield terpenoid skeletons, and then the secondary enzymatic modifications to the skeletons give rise to the functional properties and great chemical diversity of this family of natural products. The processes from IPP to terpenoid skeletons have been extensively studied but knowledge about modification steps are relatively poor. The processes of biosynthesis of terpenoids are regulated by two broad factors: the spatial and the temporal.

The formation of IPP in the plant involves two independent pathways located in separate subcellular compartments. In cytosol, IPP is derived from the long-known mevalonic acid (MVA) pathway that starts with the condensation of acetyl-CoA (Newman and Chappell, 1999). In plastids, IPP is formed from pyruvate and glyceraldehydes 3-phosphate, which is called MEP pathway, named after the key intermediate methylerythritol phosphate (Lichtenthaler, 1999). The cytosolic IPP may serve as a precursor of FPP for sesquiterpenes and triterpenes, whilst the plastidial IPP provides the precursors for GPP and GGPP for mono-, di-, and tetra-terpenes. However, cross-talk between these two IPP generation pathways is prevalent (Dudareva *et al.*, 2005), particularly in going from direction from plastids to cytosol. Then, the prenyltransferase enzymes generate GPP, FPP and GGPP from IPP in head-to-tail condensation reactions. Squalene, the direct precursor of triterpene, is formed by a head-to-head condensation reaction of two molecules of FPP, catalyzed by squalene synthase. And phytoene, the precursor of tetraterpenes, is formed by two molecules of GGPP in a manner analogous to that of squalene, which is catalyzed by phytoene synthase. These enzymes function at the branch point of terpenoid metabolism, thus playing a regulatory role in controlling IPP flux into different families of terpenoids.

The allylic prenyldiphosphaes of GPP, FPP and GGPP are used by terpene synthases (TPSs) to form mono-, sequi-, and diterpenes, respectively. After the construction of the basic parent skeletons produced by the TPSs, subsequent modifications including oxidation, reduction, isomerization, and conjugation reactions impact functional properties to the terpenoid molecules. Among these modifications, the hydroxylations or epoxidations involved in introducing oxygen atoms into the terpenoid skeletons have been extensively investigated. These reactions are performed by cytochrome P450 enzyme systems, in which the P450 monooxygenase plays a key role. Several important terpenoids biosynthesis pathways have been thoroughly studied. For example, *Artemisia annus* contains abundant terpenes, of which artemisinin is a sesquiterpene lactone with excellent antimalarial activity. Synthesis of amorpha-4, 11-diene from IPP is the first step in artemisinin biosynthesis. The amorpha-4, 11-diene synthase was cloned, expressed from *Artemisia annua* (Bouwmeester *et al.*, 1999; Wallaart *et al.*, 2001; Picaud *et al.*, 2005). Eight cDNAs encoding terpenes synthase have been isolated from *Artemisia annus*, in which four enzymes are characterized, including (−)-β-pinene synthase and β-caryophyllene synthase (Jia *et al.*, 1999; Cai *et al.*, 2002; Lu *et al.*, 2002).

Gossypol, a Malvaceae specific sesquiterpene aldehyde, is synthesized by the condensation of two moleculars of hemigossypol, which is the product of the modification of sesquiterpene (+)-δ-cadinene. One P450 monooxygenase, CYP706B1, was identified to be (+)-δ-cadinene-8-hydroxylase, and the product 8-hydroxyl-(+)-δ-cadinene is then converted to gossypol derivates (Luo *et al.*, 2001).

2.2 Biosynthesis of Phenylpropanoids

The biosynthesis pathways of lignans, lignins and flavonoids are perhaps the most studied of plant secondary metabolisms. The direct precursors of lignans and lignins are monolignols derived out of Phe and Tyr; whilst the precursor molecules of flavonoids are Phe, Tyr and malonly CoA. Phenylalanine ammonia-lyase (PAL) catalyzes the conversion of Phe to cinnamic acid (Koukol and Conn, 1961) and Tyr to *p*-coumaric acid (Neish, 1961), the first step in the phenylpropanoid pathway. In dicot Phe is the highly preferred substrate, but in monocot both Phe and Tyr could be utilized. In some plants, PAL appears to be encoded by a single gene, whereas in others it is the product of a multigene family. The ammonium ion liberated by the PAL reaction is recycled by way of glutamine synthetase and glutamate synthetase. Cinnamate-4-hydroxylase (C4H), functioning in aromatic ring hydroxylation, is an oxygen-requiring, NADPH-dependent, cytochrome P450 enzyme that catalyzes the regiospecific hydroxylation at the *para*-positon of cinnamic acid to give *p*-coumaric acid (Russell and Conn, 1967). *O*-Methyltransferases, catalyzing the transformation of a methyl group into the meta-position, uses S-adenosylmethionine (SAM) as a donor (Finkle and Nelson, 1963), whereas CoA ligation requires ATP and CoASH. This two-step ligation first generates the AMP derivative, and then converts it into the corresponding CoA ester, and two sequential NADPH-dependent reductions produce the monolignols. The monolignols are primarily converted into lignans and lignins, the first of which requires a dirigent protein to orient the putative free radical substrates in such a way that random coupling cannot occur (Davin *et al.*, 1997). However, confusion remains on whether the biosynthesis of lignins requires enzymes. We also do not yet fully understand the metabolic flux and compartmentalization of the phenylpropanoid pathway.

The flavonoid pathway is branched off the phenylpropanoid pathway from *p*-coumaryl-CoA to condense three molecules of acetate-derived malonly-CoA to generate a 6-deoxychalcone, which is catalyzed by chalcone synthase (CHS). CHS is a dimeric polyketide synthase with each subunit at about 42 kDa. Then, chalcone isomerase (CHI) catalyzes a stereospecific ring closure isomerization step to form some flavanones, which is shared by most of the flavonoid biosynthesis pathways. The isomerization of the flavanones leads to the isoflavonoid branch point catalyzed by two enzymes. Isoflavone synthase, an NADPH-dependent cytochrome P450 enzyme, catalyzes the first step of a 1, 2 aryl migration and hydroxylation to give the 2-hydroxyisoflavanones. Dehydration of the 2-hydroxyisoflavanone dehydratase (IFD), forms the isoflavonoids. In general flavonoid metabolism, the second branching point involves dehydration of naringenin at the C-2/C-3 positions to give abundant flavones. This conversion is catalyzed by flavone synthase (FNS), which varies in enzymatic type depending on the plant species. The third branch point is stereospecific 3-hydroxylation of naringenin to give dihydroflavonols, which is catalyzed by flavanone 3-hydroxylase, a Fe^{2+}-requiring, α-ketoglutarate-dependent dioxygenase. The subsequent species- and tissue-specific enzymatic conversions could create vast arrays of structurally diverse groups of flavonoids. Condensed tannins are formed from flavonoids, however, the enzymology associated with those coupling processes, chain extension mechanisms, and oxidative modifications is not yet established.

3 Plant genomes for secondary metabolism

Early efforts at natural product isolation and enzyme discovery have their limits. The compounds that can be obtained by traditional isolation method, even those of trace components, are consequently not a comprehensive accounting of the organism's biosynthetic capacity, but a reflection of the state of the tissues upon harvest. As natural product biosynthesis is heavily influenced by external stimuli such as microbial infection or herbivory, the extracts of un-induced plants consequently contain only a subset of the products that these organisms can biosynthesize. The enzymes obtained by the precursor-feeding, mutant screening and cDNA expression experiments are also not thoroughly reflecting the organism's biosynthetic capacity, since the enzymes involved in secondary metabolim are often tissue-specific and regulated under environmental influences. A complete understanding of plant secondary metabolism and their biosynthesis pathway thus requires new and comprehensive methods to reveal their molecular basis and to overcome the limits mentioned above.

Publication of the first two bacterial genomes in 1995 marked the beginning of the genomic era (DellaPenna and Last, 2008). It has been proved that the genomics and bioinformatics could serve as predictors of new molecules and enzymes. The sequenced streptomycetes of *S. avermitilis* (Ōmura *et al.*, 2001) and *S. griseus* (Ohishi *et al.*, 2008) could make 2–3 natural products but harbor 25–30 predicted biosynthetic gene clusters. To date, we are missing 90% of the natural product biosynthetic capacity of even the workhorse producers. If even 20% of the 20–25 cryptic molecules were novel, the current knowledge base from streptomyctes would double (Walsh and Fischabch, 2010). Thus bacterial genomics and bioinformatics have become as important as chemistry in categorizing known natural products and identifying likely unknown variants to be discovered.

Since the genes for plant natural product pathways are rarely physically clustered as bacterial genes do, efforts to decipher plant natural product pathways have lagged behind those to bacterial pathways. However, bioinformatics is a useful tool to identify the metabolic functions of unknown plant genes once the completed genomes of the plant are obtained. The development of such tools based on plant genomics and bioinformatics as protein family-based analysis, contextual genomics approaches, cell-specific comparisons, and co-expression analysis allows genome-based research in plant metabolism to be more feasible (DellaPenna and Last, 2008). Thalianol is a novel triterpene first uncovered by heterologous expression of enzymes exploited from *Arabidopsis thaliana* genomic information and then detected in *A. thaliana* at a low level, which is a good indicator that genome mining can uncover secondary metabolites that eluded classical methodologies (Fazio *et al.*, 2004). It is estimated that in *Arabidopsis thaliana*, about 5 000 genes (about a quarter of all) are involved in secondary metabolism (Gierl and Frey, 2001). Therefore, the current public availability of draft or completed genomes for rapidly increasing numbers of organisms of different taxonomic groups creates unprecedented opportunities to study individual plant enzymes, pathways, and metabolic networks.

Investigations using genomics and bioinformatics methods have produced extensive knowledge and interpretation of the chemical diversity in plants. As discussed above, plants use limited building blocks to construct structurally diversified and complex secondary metabolites. Not surprisingly, the biosynthetic pathways of these secondary metabolites are accordingly complex (Fischbach *et al.*, 2008). The set of proteins that comprise a complete biosynthetic pathway can be twice the size of the ribosome, even though the ribosome translates thousands of different proteins, whereas the biosynthetic pathway produces a few small molecules. Furthermore, some enzymes involved in natural product biosynthesis have a broad substrate tolerance, which is now firmly supported by experimental evidence gained for all major natural product pathways (Firn and Jones, 2003). A good example of this tolerance is the multifunctional enzyme of *Arabidopis* LUP1 (At1G78970) that converts oxidosqualene to mixtures of at least 6–7 distinct triterpene alcohols. Such enzymes capable of acting on more than one substrate would be expected to facilitate branching pathways, and at the extreme, to participate in a matrix grid, which creates more chemical diversity with limited enzymes. In petunia flowers, for example, three enzymes (F3H, F3′5′H and F3′H) can produce five different products (eriodictyol, pentahydroxyflanone, dihydromyricetin, dihydroquercetic anddihydrokaempferol) from naringenin (Holton and Cornish, 1989).

Also, genes functioning in secondary metabolism are generally more divergent than those coding for proteins involved in primary metabolism. For example, in *Arabidopsis thaliana*, a family of 40 terpenoid synthase genes (*AtTPS*) was discovered by genome sequence analysis. Among them, thirty-two *AtTPS* genes are attributed as putative monoterpene synthases, sesquiterpene synthases or diterpene synthases of secondary metabolism. In contrast, only two *AtTPS* genes have known functions in hormone metabolism, namely gibberellin biosynthesis (Aubourg *et al.*, 2002). This striking difference in rates of gene diversification in primary (hormone) and secondary metabolisms is relevant for an understanding of the evolution of natural product diversity. Alignment of the amino acid sequences of plant terpene syntheses divided the TPS family into seven subfamilies, designated TPSa to TPSg. Each subfamily has a minimum of 40% sequence identity among members and has, in general, similar functions. However, specific product profiles of members of the same subfamily can be quite diverse and cannot be predicted based on sequence alone. There are two strategies used in terpenoids synthesis to create chemical diversity. One is that different terpene synthases use the same substrate to produce different products, and the other is one terpene synthase produces multiple products.

Plants could also mix genes to create chemical diversity by juxtaposing distinct but chemically compatible biosynthetic systems. Tailored natural products are the most important results of this strategy (Walsh and Fischabch, 2010), which are produced by functionalizing the core scaffold, often occurring later in the pathway. Tailoring enzyme chemistries can be grouped into two broad categories: group transfer reactions and oxidative transformations. Nearly all group transfers involve coupling an electrophilic fragment of a cosubstrate or primary metabolite to a nucleophilic N, O, or S in the natural product skeleton. These cosubstrates include NDP-sugars (such as UDP-Glucose) as glycosyl donors, S-adenosylmethionine as methyl donors, acyl-CoA as an acyl donor and the correponding enzymes are glycosyltransferases, methyltransferases, and acyltransferases, respectively. The so-called BAHD superfamily enzymes are a large group of plant-specific acyl-CoA dependent *O*-or *N*-acyltransferases identified recently. Most of them fall into two functional families: alcohol acetyltransferases responsible for forming aroma/flavor volatile acetate esters, such as geranyl acetate, phenylethyl acetate, and benzyl acetate (Dudareva and Pichersky, 2000) and, anthocyanin/flavonoid acyltransferases, primarily malonyltranserases and hydroxycinnamoyltransferases, responsible for modifying polyphenolics (Yu *et al.*, 2008). Several BAHD members synthesize and modify a variety of other metabolites, such as shikimate-phenylpropanoid-derived phytoalexins, alkaloids, terpenoids, and polyamines (Luo *et al.*, 2009). The sequences of BAHD genes are highly divergent, showing only 10–30% similarity at the amino-acid level, consistent with their functional diversity (D'Auria, 2006). Using the conserved sequence motifs HXXXD and DFGWG of BAHD members, 94 and 61 putative BAHD genes were identified from the genome sequences of *Populus* and *Arabidopsis*, respectively (Yu *et al.*, 2009). The ratio of the number of BAHD genes in *Populus* to that of *Arabidopsis* is consistent with the estimation of the entire genome. Gene chromosomal distribution demonstrates that both individual gene tandem duplication and large-scale genome segmental duplication events appear to have exclusively contributed to the current complexity of the BAHD gene superfamilies in both *Populus* and *Arabidopsis*. Having a greater number of diverse BAHD enzymes meets the needs of the biosynthesis and modification of a repertoire of the secondary metabolites in the plants.

Oxidative transformations may plan an important role in creating chemical diversity in plants. Terpenoid pathways generate product with dramatic structural diversities from simple building blocks by first generating a reduced, unreactive polycyclic intermediate and then tailoring it with oxygen-based functionality (Walsh and Fischabch, 2010). Unlike group transfer reactions that simply introduce a structural fragment in the core skeleton, oxidative transformations can also generate a novel skeleton through structural rearrangement as the oxygen-based functionality serves as a new reaction centre. In the flavonoid, lignan, and lignin pathway, oxidative transformations also catalyze aromatic and aliphatic hydroxylations, and skeleton formation (Ayabe and Akashi, 2006). In plants, heme iron-containing enzymes of the P450 monooxygenase superfamily are widespread oxidative tailoring enzymes of natural products, adding thousands of genes falling into 126 families and 464 subfamilies. Plant genome encodes more P450 enzymes than other organisms. For example, there are 246 genes in *Arabidopsis thaliana*, 356 in rice, 312 in poplar, and 457 in grape, and the number of P450 genes is estimated at up to 1% of total gene annotations of each plant species (Mizutani and Ohta, 2010); whereas there are 57 P450 genes in *Homo sapiens*, 105 in *Mus musculus* (Nelson *et al.*, 2004), 86 in *Drosophila melanogaster* (Chung *et al.*, 2008), and 80 in *Caenorhabditis elegans* (Menzel *et al.*, 2001). This observation is in accordance with the diversified reactions P450 enzymes catalyzed in plant secondary metabolism pathway such as hydroxylation, epoxidation, dealkylation, isomerization, dehydration, carbon-carbon cleavage, decarboxylation, nitrogen and sulfur oxidation, dehalogenation, and deamination. P450 diversification during evolution was one of the primary driving forces of phytochemical diversity.

Interestingly, P450 enzymes can form gene cluster for biosynthesis of secondary metabolites. For instance, a triterpene gene cluster (Field and Osbourn, 2008), three diterpene gene clusters (Shimura *et al.*, 2007; Swaminathan *et al.*, 2009), and a benzoxazinoid gene cluster (Gierl and Frey, 2001) were found in *Arabidopsis*, rice and maize, respectively (Table 1). Clustering in plants facilitates the inheritance of beneficial combinations of genes, and avoids the accumulation of toxic intermediates.

The production of secondary metabolites by plants is influenced by developmental, environmental, pathogen and symbiont signals, which leave the investigation of secondary metabolites pathway even more complicated. It is now generally believed that, in plants, transcription factors play a major role in the regulation of secondary metabolism pathways, as many other aspects of plant growth and development. The phenylpropanoid pathway has been the leading model for studies on plant gene regulation but little is known about any of the other major metabolic pathways (Davies and Schwinn, 2003). Some transcription factors of flavonoid pathways have been cloned, such as C1, PAR1, ANTHOCYANIN2 (MYB); R, ANTHOCYANIN1, TT8 (Bhlh), etc. Interestingly, some of the anthocyanin regulatory genes also impact other processes. Rosea1 of *Antirrhinum* and An1, An2 and An11 of *Petunia* also regulate vacuolar pH, and additionally, An1 influences seed coat epidermal cell development (Davies and Schwinn, 2003). The best model for studies of such overlapping pathways is *Arabidopsis*, in which the regulatory pathways for anthocyanin and proanthocyanidin production share components with those controlling developmental processes such as trichome formation, root epidermal cell development and seed mucilage production (Davies and Schwinn, 2003). Transcription factors can be the main target of the secondary metabolic engineering. As the activity of the biosynthetic genes of secondary metabolism appears to be determined primarily by the expression patterns of the regulatory genes, altering the patterns of regulatory gene expression may allow the temporal and spatial modification of secondary metabolite production. For example, in apple MYB10 is an anthocyanin-regulating transcription factor; an allelic rearrangement in the gene promoter of *MYB10* has generated an autoregulatory locus, which is sufficient to account for the increase in MYB10 transcript levels and subsequent ectopic accumulation of anthocyanins in the plant, leading to a striking phenotype that includes red foliage and red fruit flesh (Espley *et al.*, 2009). Another example is the GaWRKY1, a WRKY transcription factor in cotton that regulates gossypol content and expression of biosynthesis pathway gene (+)-δ-cadinene synthase, by binding to the promoter and activates its spacial and temporal expression. (Xu, 2004) Therefore, the identification of defined transcription factor genes provides tools for modulating both the amount and distribution of secondary metabolites in plants.

Table 1 Example of plant P450 clusters identified functionally

P450s name	Other genes name	Products	Species
At5g48000	*At5g48010*	Desaturated thalina-diol	Arabidopsis
At5g47990	*At5g47980*	(triterpenoid)	
CYP99A2	AK103462	Momilactone A	Rice
CYP99A3	*OsCycl*	(diterpenoid)	
	OsKS4		
Cyp71Z6	*Os-CPS2*	Phytocassanes A-E	Rice
Cyp71Z7	*Os-KSL7*	(diterpenoid)	
Cyp76M5	*Os-KSL5*		
Cyp76M6	*Os-KSL6*		
Cyp76M7			
Cyp76M8			
CYP99A2	*Os-CPS4*	Momilactone A&B	Rice
CYP99A3	*Os-KSL4*	(diterpenoid)	
	Os-MAS		
BX2		DIBOA	Maize
BX3		(benzoxazinone)	
BX4			
BX5			

In addition to resolving the puzzle of how structurally diversified metabolites are synthesized in plants, the genomics and bioinformatics methods can explain how and why these compounds occurred and evolved and how new compounds are produced. It seems most reasonable to assume that the precursors and pathways for the generation of natural products arose from mutations of enzymes involved in the synthesis of primary metabolites. The large-scale genome segmental duplication and individual gene tandem duplication of primary pathway genes allowed the original enzymatic function retained in the plant, while new functions evolve in the enzyme encoded by the duplicate genes under the pressure of natural selection, which can generate secondary metabolites. This may explain why primary and secondary metabolites cannot readily be distinguished on the basis of precursor molecules, chemical structures, or biosynthetic origins. Similarly, duplication of secondary pathway genes followed by divergence can form new metabolites once a secondary pathway has been branched from the primary pathway and this appears to have been the most common means for diversified secondary metabolites evolution in plants (Gang, 2005). New metabolites may also arise due to loss of enzymatic activity, which can occur due to loss of gene expression. The intermediate molecules in the affected pathway may build up to levels that were not previously present, enabling other enzymes to act on these metabolites. Alternatively, changes in regulatory gene expression may not necessarily lead to complete loss of enzyme activity from the plant, but may cause the enzyme to be produced in a different cell, tissue, or organ type (Gang, 2005). As a result, the enzyme can work on a different substrate to produce new metabolites. If these products are favored by natural selection, then new chemical diversity is generated.

4 Perspectives

There are more than 400 families, more than 10 000 genera and nearly 300 000 species of angiosperms on earth. Most of the secondary metabolic pathways are, although to a different extent, taxa-specific. The great richness of plant species is a huge treasure for plant secondary metabolites, leaving the enormous biosynthetic potential of plant cells to be exploited. Recent advances in metabolomics, and other areas of study, have made it possible to reveal the dynamic changes of many types of secondary metabolites in a plant sample. The updated knowledge in metabolomics has helped genetic engineering in introducing whole pathways to produce medicinally valuable products in organisms that lack them. The genetic engineering artemisin biosynthesis pathway in yeast and resulting in the accumulation of artemisinic acid is a pioneering success of synthetic biology. Although the development of plant genomics and functional genomics has gained much progress, the genetic maps of biosynthetic pathways are still far from complete, and the networks regulating different pathways globally are still poorly understood. As plant metabolites are of great interests and importance to human health, we expect to see more efforts into research and engineering of plant secondary metabolism.

References:

Aubourg S, Lecharny A, Bohlmann J, 2002. Genomic analysis of the terpenoid synthase (AtTPS) gene family of *Arabidopsis thaliana* [J]. *Molecular Genetics & Genomics*, **267**: 730—745

Ayabe SI, Akashi T, 2006. Cytochrome P450s in flavonoid metabolism [J]. *Phytochemistry Reviews* , **5**: 271—282

Bouwmeester HJ, Wallaart TE, Janssen MHA *et al.*, 1999. Amorpha-4, 11-diene synthase catalyses the first probable step in artemisinin biosynthesis [J]. *Phytochemistry*, **52**: 843—854

Buchanan BB, Gruissem W, Jones RL, 2000. Biochemistry and Molecular Biology of Plants [M]. Rockville: American Society of Plant Physiologists

Cai Y, Jia JW, Crock J *et al.*, 2002. A cDNA clone for β-caryophyllene synthase from *Artemisia annua* [J]. *Phytochemistry*, **61**: 523—529

Chen F, Liu CJ, Tschaplinske TJ *et al.*, 2009. Genomics of secondary metabolism in *Populus*: interactions with biotic and abiotic environments [J]. *Critical Reviews in Plant Sciences*, **28**: 375—392

Chung H, Sztal T, Pasricha S *et al.*, 2008. Characterization of Drosophila melanogaster cytochrome P450 genes [J]. *Proceedings of the National Academy of Sciences of the United States of America*, **106**: 5731—5736

Connolly JD, Hill RA, 2010. Triterpenoids [J]. *Natural Product Reports*, **27**: 79—132

D'Auria JC, 2006. Acyltransferases in plants: a good time to be BAHD [J]. *Current Opinion in Plant Biology*, **9**: 331—340

Davies KM, Schwinn KE, 2003. Transcriptional regulation of secondary metabolism [J]. *Functional Plant Biology*, **30**: 913—925.

Davin LB, Wang HB, Crowell AL *et al.*, 1997. Stereoselective bimolecular phenoxy tadical coupling by an auxiliary (dirigent) protein without an active center [J]. *Science*, **275**: 362—366

DellaPenna D, Last RL, 2008. Genome-enabled approaches shed new light on plant metabolism [J]. *Science*, **320**: 479—481

Dewick PM, 2002. Medicinal Natural Products [M]. Baffins lane: John Wiley & Sons Ltd

Dey PM, Harborne JB, 1997. Plant Biochemistry [M]. San Diego: Academic Press Ltd

Dudareva N, Andersson S, Orlova I *et al.*, 2005. The nonmevalonate pathway supports both monterpene and sesquiterpene formation in snapdragon flowers [J]. *Proceedings of the National Academy of Sciences of the United States of America*, **102**: 933—938

Dudareva N, Pichersky E, 2000. Biochemical and molecular genetic aspects of floral scents [J]. *Plant Physiolosy*, **122**: 627—633

Espley RV, Brendolise C, Chagné D *et al.*, 2009. Multiple repeats of a promoter segment causes transcription factor autoregulation in red apples [J]. *The Plant Cell*, **21**: 168—183

Fazio GC, Xu R, Matsuda SPT, 2004. Genome mining to identify new plant triterpenoids [J]. *Journal of the American Chemical Society*, **126**: 5678—5679

Field B, Osbourn AE, 2008. Metabolic diversification-independent assembly of operon-like gene clusters in different plants [J]. *Science*, **320**: 543—547

Fischbach MA, Walsh CT, Clardy J, 2008. The evolution of gene collectives: how natural selection drives chemical innovation

［J］．*Proceedings of the National Academy of Sciences of the United States of America*，**105**：4601—4608

Firn RD，Jones CG，2003．Natural products-a simple model to explain chemical diversity ［J］．*Natural Product Reports*，**20**：382—391

Finkle BJ，Nelson RF，1963．Enzyme reactions with phenolic compounds：a meta-O-methyltransferase in plants ［J］．*Biochimica et Biophysica Acta*，**78**：747—749

Fraga BM，1997．Natural sesquiterpenoids ［J］．*Natural Product Reports*，**14**：145—162

Gang DR，2005．Evolution of flavors and scents ［J］．*Annual Review of Plant Biology*，**56**：301—325

GierlA，Frey M，2001．Evolution of benzoxazinone biosynthesis and indole production in maize ［J］．*Planta*，**213**：493—498

Grayson DH，2000．Monoterpenoids ［J］．*Natural Product Reports*，**17**：385—419

Hanson JR，2004．Diterpenoids ［J］．*Natural Product Reports*，**21**：785—793

Holton TA，Cornish EC，1989．Genetics and biochemistry of anthocyanin biosynthesis ［J］．*The Plant Cell*，**7**：1071—1083

Jia JW，Crock J，Lu S *et al*．，1999．（3R）-Linalool synthase from Artemisia annua L．：cDNA isolation，characterization，and wound induction ［J］．*Archives of Biochemistry and Biophysics*，**372**：143—149

Koukol J，Conn EE，1961．The metabolism of aromatic compounds in higher plant：purification and properties of the phenylalanine deaminase of *Hordeum vulgare* ［J］．*The Journal of Biological Chemistry*，**236**：2692—2698

Lichtenthaler HK，1999．The 1-deoxy-*D*-xylulose-5-phosphate pathway of isoprenoid biosynthesis in plants ［J］．*Annual Review of Plant Physiology and Plant Molecular Biology*，**50**：47—65

Lu S，Xu R，Jia JW *et al*．，2002．Cloning and functional characterization of a β-pinene synthase from Artemisia annua that shows a circadian pattern of expression ［J］．*Plant Physiolosy*，**130**：477—486

Luo J，Fuell C，Parr A *et al*．，2009．A novel polyamine acyltransferase responsible for the accumulation of spermidine conjugates in *Arabidopsis* seed ［J］．*The Plant Cell*，**21**：318—333

Luo P，Wang YH，Wang GD *et al*．，2001．Molecular cloning and functional identification of（+）-d−cadinene-8-hydroxylase，a cytochrome P450 monooxygenase（CYP706B1）of cotton sesquiterpene biosynthesis ［J］．*The Plant Journal*，**28**：95—104

Menzel R，Bogaert T，Achazi R，2001．Asystematic gene expression screen of *Caenorhabditis elegans* cytochrome P450 genes reveals CYP35 as strongly Xenobiotic inducible ［J］．*Archives of Biochemistry and Biophysics*，**395**：158—168

Mizutani M，Ohta D，2010．Diversification of P450 genes during land plant evolution ［J］．*Annual Review of Plant Biology*，**61**：1—25

Neish AC，1961．Formation of *m*- and *p*-coumaric acids by enzymatic deamination of the corresponding isomers of tyrosine ［J］．*Phytochemistry*，**1**：1—24

Nelson DR，Zeldin DC，Hoffman SMG *et al*．，2004．Comparison of cytochrome P450（CYP）genes from mouse and human genomes，including nomenclature recommendations for genes，pseudogenes and alternativesplice variants ［J］．*Pharmacogenetics and Genomics*，**14**：1—18

Newman JD，Chappell J，1999．Isoprenoid biosynthesis in plants：carbon partitioning within the cytoplasmic pathway ［J］．*Critical Reviews in Biochemistry and Molecular Biology*，**34**：95—106

Ohishi Y，Ishikawa J，Hara H *et al*．，2008．Genome sequence of the streptomycin-producing microorganism *Streptomyces griseus* IFO 13350 ［J］．*The Journal of Bacteriology*，**190**：4050—4060

Ōmura S，Ikeda H，Ishikawa J *et al*．，2001．Genome sequence of an industrial microoganism *Streptomyces avermitilis* deducing the ability of producing secondary metabolites ［J］．*Proceedings of the National Academy of Sciences of the United States of America*，**98**：12215—12220

Picaud S，Olofsson L，Brodelius M *et al*．，2005．Expression，purification，and characterization of recombinant amorpha-4，11-diene synthase from Artemisia annua L． ［J］．*Archives of Biochemistry and Biophysics*，**436**：215—226

Russell DW，Conn EE，1967．The cinnamic acid 4-hydraxylase of pea seedlings ［J］．*Archives of Biochemistry and Biophysics*，**122**：256—258

Shimura K，Okada A，Okada K *et al*．，2007．Identification of a biosynthetic gene cluster in rice for momilactones ［J］．*The Journal of Biological Chemistry*，**282**：34013—34018

Swaminathan S，Morrone D，Wang Q *et al*．，2009．CYP76M7 is an *ent*-Cassadiene C11α-Hydroxylase defining a second multifunctional diterpenoid biosynthetic gene cluster in rice ［J］．*The Plant Cell*，**21**：3315—3325

Trethewey RN，2004．Metabolite profiling as an aid to metabolic engineering in plants ［J］．*Current Opinion in Plant Biology*，**7**：196—201

Wallaart TE，Bouwmeester HJ，Hille J *et al*．，2001．Amorpha-4，11-diene synthase：cloning and functional expression of a key enzyme in the biosynthetic pathway of the novel antimalarial drug artemisinin ［J］．*Planta*，**212**：460—465

Walsh CT，Fischabch MA，2010．Natural products version 2．0：connecting genes to molecules ［J］．*Journal of the American Chemical Society*，**132**：2469—2493

Yu XH，Gou JY，Liu CJ，2009．BAHD superfamily of acyl-CoA dependent acyltransferases in *Populus* and *Arabidopsis*：bioinformatics and gene expression ［J］．*Plant Molecular Biology*，**70**：421—442

Yu XH，Chen MH，Liu CJ，2008．Nucleocytoplasmic-localized acyltransferases catalyze the malonylation of 7-*O*-glycosidic（iso）flavones in *Medicago truncatula* ［J］．*The Plant Journal*，**55**：382—396

Ziegler J，Facchini PJ，2008．Alkaloid biosynthesis：metabolism and trafficking ［J］．*Annual Review of Plant Biology*，**59**：735—769