CellPress
REVIEWS

Forum

Identifying Missing Biosynthesis Enzymes of Plant Natural Products

Thomas Dugé de Bernonville,[1]
Nicolas Papon,[2]
Marc Clastre,[1]
Sarah E. O'Connor,[3,*] and
Vincent Courdavault[1,*]

Elucidating plant-specialized biosynthetic pathways has always constituted a laborious task, notably for natural products with high pharmaceutical values. Here, we discuss emerging omics-based strategies that facilitate the identification of genes from these complex metabolic pathways, paving the way to engineered supplies of these compounds through synthetic biology approaches.

Finding a Needle in a Haystack

Plants represent a remarkable source of natural metabolites, many of which remain the basis of the pharmacopoeia used by humans to treat various disorders [1]. Highly potent compounds are found within monoterpene indole alkaloids of the Apocynaceae plant family, yew taxane-type terpenoids, mayapple lignans, poppy isoquinoline alkaloids, and hemp cannabinoids. Due to their pharmaceutical importance, the biosynthetic pathways responsible for the production of these compounds *in planta* have attracted the attention of many research groups for decades. This area remains in focus as it is essential to be able to propose cheaper and high-throughput alternatives of production of these valuable pharmaceutical compounds, in a short period.

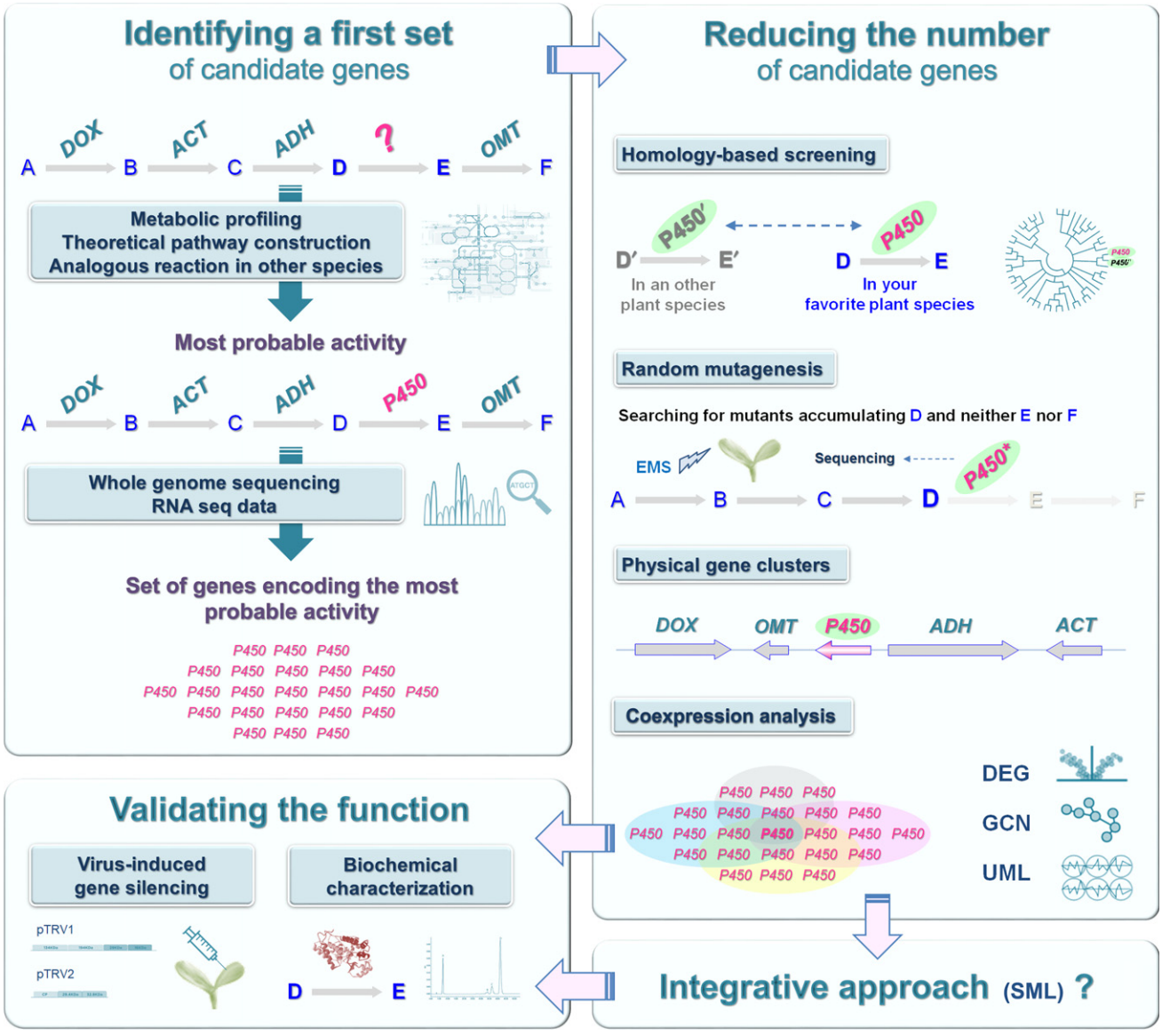
While most of the enzymes catalyzing biosynthesis steps have been progressively identified using **sequence homology** (see Glossary)-based cloning or protein purification, these discovery efforts have always been laborious, particularly until the arrival of massive omics resources such as **next generation sequencing (NGS)**. This is likely because downstream branches of these pathways: (i) differ among plant species, even within the same genus; and (ii) involve enzymes from superfamilies containing several tens to hundreds of members [2]. Such superfamilies include reductases, oxidases, methyltransferases, and cytochrome P450s [3]. Cytochrome P450s in particular are key components of plant biosynthetic pathways and may catalyze multiple types of reactions, including hydroxylation, epoxidation, demethylation, dealkylation, decarboxylation, and C–C bond cleavage [4]. Fortunately, recent advances in omics-based strategies have accelerated the identification of missing biosynthetic enzymes. Below, we briefly highlight these advances, also summarized in Figure 1.

Identifying First Sets of Candidate Genes through Metabolomics and Transcriptomics

Identifying enzymes catalyzing steps of given pathways requires a prior knowledge on potential reaction scenarios leading to the formation of the expected compound, as illustrated on the left part of Figure 1. Metabolomics are typically used to identify preferential accumulation sites of specialized metabolites and to detect plausible reaction intermediates. Molecular networking also represents a considerable advance in the processing of the resulting metabolomics data acquired with either liquid chromatography or gas chromatography coupled to mass spectrometry by connecting related separated compounds based on their mass spectra [5]. Ideally, connected compounds may reflect an entire biosynthetic branch. Once a reaction sequence has been postulated, a type of enzymatic activity required for each step may be hypothesized based on literature data or bioinspired organic chemistry. Activity of enzymes relies on specific functional conserved domains that can drive their identification. For instance, a hydroxylation step identified as described above may be catalyzed by a cytochrome P450 or a dioxygenase, each bearing their own functional domain. Defining a first set of candidate genes requires access to genomes or transcriptome assemblies, providing gene or transcript sequences, respectively, to look for such specific domains. Although challenging in terms of computational requirements, whole genomes may be sequenced at lower cost from short DNA reads generated through NGS. **RNA-seq**-based

Glossary

Next generation sequencing (NGS): second generation technologies refers to the high-throughput sequencing of DNA fragment ends, usually outputting millions of short reads (up to 250 nt long). The third generation technologies sequence single DNA molecules, generating long reads (at least >1000 nt long).
RNA-seq: high-throughput sequencing of reverse-transcribed mRNA. Short read-based RNA-seq is typically a two-step workflow, including: (i) transcript reconstruction (genome-based or *de novo*); and (ii) gene expression inference from read attribution to reconstructed transcript sequences.
Self-organizing maps: an unsupervised machine learning algorithm used to cluster genes into a given number of nodes forming a 2D map. In this map, regions contain nodes reflecting genes with similar expression levels.
Sequence homology: homology defines similarity between two nucleic or protein sequences. Similarities can be searched against specific databases with the well-known BLAST algorithm. Phylogenies compare a set of sequences to highlight evolutionary relationships through statistical approaches.
Virus-induced gene silencing (VIGS): a reverse genetics approach allowing transient plant gene downregulation mediated by the infection of a modified virus harboring a short sequence targeting a specific gene.
Whole genome assembly: a complex task due to inherent genome complexity and requires a combination of short reads and long reads to ensure the reconstruction of long gDNA scaffolds.



Trends in Pharmacological Sciences

Figure 1. Workflow for Identification of Missing Steps in Plant-Specialized Metabolic Pathways. Identification process starts from a thorough understanding of possible chemical reactions transforming substrate A to product F with biosynthetic intermediates B, C, D, and E. Conversions may involve already characterized reactions, here, for example, those involving a dioxygenase (DOX), a acetyltransferase (ACT), an alcohol dehydrogenase (ADH), and an O-methyltransferase (OMT). In this example, we illustrate identification of the enzymatic step converting compound D to E. Previous data show that E is a hydroxylated form of D and, based on literature precedence, it is expected that this hydroxylation is catalyzed by a cytochrome P450 monooxygenase. The first step consists of the identification of genes encoding cytochrome P450s in predicted gene sets from a genome (whole-genome sequencing) or a transcriptome (RNA-seq) assembly. This will result in a large list of candidate cytochrome P450s as it is a large family in plants. The second step aims at reducing this large list through different approaches described in the main text. This includes homology-based, random mutagenesis, physical gene clusters in genomes and coexpression analysis. This last analysis may be performed using either differentially expressed genes (DEG), gene coexpression networks (GCN), or unsupervised machine learning (UML). The resulting candidates are then functionally validated (pTRV1 and 2 represent plasmids encoding the two genomic components of the tobacco rattle virus classically used for virus induced gene silencing). An integrative approach such as supervised machine learning (SML) will merge the power of each approach to refine candidate genes. An SML approach should be able to integrate different variables (gene expression, gene clusters, etc.) and set rules to correctly class genes in a given metabolic pathway.

transcriptome analysis is simpler to conduct than **whole genome assembly** because the transcriptome only covers a small part of the genome. Based on genome or transcriptome annotation, functional domains can be systematically attributed to define first sets of candidate genes that will require a further prioritization through the methodological aspects described below.

Prioritization by Homology-Based Screening

One of the most direct and obvious ways to unravel missing steps is to take inspiration from previous works reported in the literature and screen for homology to orthologous gene sequences of already characterized biosynthesis steps in other plant species (Figure 1). This can be done by searching homologies [e.g., with basic local alignment search tool (BLAST)] or constructing phylogenies statistically fitted on evolutionary models encompassing protein sequences from multiple species [6]. For example, Farrow *et al*. [7] recently identified the ibogamine 10-hydroxylase from the African shrub (*Tabernanthe iboga*) by searching for genes with homology to tabersonine 16-hydroxylase from Madagascar periwinkle (*Catharanthus roseus*), which catalyzes indole ring hydroxylation at the same position. However, while often successful, this approach must be considered with caution since it has been demonstrated in the above example that the closest ortholog does not necessarily catalyze a similar reaction [7].

Prioritization by Random Mutagenesis

Unlike the above described targeted approach, wide-scale untargeted studies can be deployed using ethyl methanesulfonate-based random mutagenesis (Figure 1) [8]. In this approach, mutagenized plants are screened for changes in the production of the desired metabolite. Plants with altered profiles are then sequenced to identify the source of the mutation by NGS that is responsible for changes in the metabolism. This is a labor-intensive but powerful strategy already applied to *C. roseus* to isolate mutants with altered alkaloid contents. This approach led to the identification of an O-acetylstemmadenine oxidase catalyzing a major oxidation step in catharanthine and tabersonine biosynthesis [9].

Prioritization by Searching Physical Gene Clusters

Since plant biosynthetic pathways sometimes involve enzymes encoded by genes physically clustered in the same genomic region, genome sequence may improve homology searches by the analysis of genomic neighbors. Candidate genes can thus be refined through the *in silico* processing of large sequencing data using previously characterized biosynthesis genes as baits (Figure 1). It is particularly efficient with more complete genome sequences composed of very long DNA sequence assemblies, as recently exemplified with the identification of a gene cluster encoding five enzymes involved in thebaine biosynthesis in opium poppy (*Papaver somniferum*) [10]. With the explosion in the availability of plant genome sequences in the coming years, new gene clusters are likely to be discovered. However, one of the disadvantages of this method is that biosynthetic pathways show only few, if no, physically clustered genes, such as recently observed in happy tree (*Camptotheca acuminata*) [11].

Prioritization by Gene Coexpression Analysis

Candidate gene selection may also rely on expression pattern similarities among tissues or experimental conditions displayed by genes related to a common biosynthetic branch (Figure 1). The underlying hypothesis is that metabolites accumulated in a tissue-specific manner should be produced through a chain of enzymes displaying a similar spatio-temporal expression pattern, which should be visible at the transcript level.

Such coexpression can guide identification of missing steps by generating gene lists corresponding to groups of genes that have similar expression across tissues. Groups of similarly expressed transcripts that contain genes already identified as part of the investigated metabolic pathway are used to identify candidate genes for the missing steps. It typically requires the measurement of genome-wide gene expression in different tissues or experimental conditions using RNA-seq [12]. Three major approaches can be used on the generated data (expression matrices containing genes in rows and sample types in columns) for coexpression analyses and eventual identification of missing biosynthetic enzymes.

A first approach entails the statistical comparison of contrasted samples (e.g., roots versus leaves or control versus treatment) to identify differentially expressed genes. In this way, geissoschizine oxidase was found to be upregulated in stressed leaves of *C. roseus*, together with strictosidine glucosidase that commonly act in the monoterpene indole alkaloid biosynthetic pathway [13]. More extensively, gene expression profiles can be compared by a hierarchical clustering approach in multiple sample comparisons, such as performed to identify enzymes converting matairesinol into etoposide aglycone in the mayapple (*Podophyllum hexandrum*) lignan biosynthetic pathway [14].

A second method is based on the construction of gene coexpression networks (GCNs) to provide a more exhaustive view of coexpression relationships among genes [12]. GCNs visualize similarities in gene expression profiles where distances between each possible gene pairs are calculated using specific metrics (e.g., Pearson correlation coefficient) and only the best coexpressed gene

pairs are retained to construct a GCN. In GCN type representations, genes (also referred to as 'nodes') are connected by edges representing these distances [12]. This approach has been successfully used to streamline the identification of precondylocarpine acetate synthase and tabersonine synthase in *C. roseus* [15].

Another method for analyzing gene coexpression is based on unsupervised machine learning (UML). UML methods group similarly expressed genes into a specific expression cluster. These methods are unsupervised because they cluster genes according to their expression levels without prior knowledge of their function. UML has been successfully used several times, in particular with **self-organizing maps**. As an example, this algorithm has been used to identify the sarpagan-bridge enzyme involved in ajmaline biosynthesis in the devil pepper (*Rauvolfia serpentina*), an Apocynaceae closely related to *C. roseus* [16].

Validating Functions of Candidate Genes

Elucidating candidate gene functions is a mandatory step but can be laborious and time consuming especially in nonmodel plant species for which no mutant libraries or no quantitative genetics data are available, albeit they are the main source of specialized metabolites. While biochemical characterization relying on recombinant protein expression can be envisaged for a small number of candidates, it always requires access to potential enzyme substrates. Over the last few years, wider gene function analyses have been enabled through the development of efficient and straightforward reverse genetic approaches based on transient transcript degradation through **virus-induced gene silencing (VIGS)**, notably [17]. Besides confirming involvement in biosynthetic pathways, such an approach can also provide evidence of gene function through the identification of accumulated biosynthetic intermediates resulting from silencing [13]. However, final biochemical characterization is still required to confirm VIGS results. Finally, for validation of multiples genes from the same pathway, heterologous reconstitution of partial pathways can be performed by simultaneous gene co-overexpression combined to biosynthetic precursor feeding, as performed for podophyllotoxin and tabersonine pathways [14,15].

Concluding Remarks and Future Perspectives

The present forum article briefly describes current procedures used to characterize missing steps from plant metabolic pathways. The advent of NGS technologies have largely fueled these procedures, allowing completion of pathway knowledge in several medicinal plants.

According to the different strategies depicted in Figure 1 and described above, candidate genes are selected using their sequence and/or expression properties. However, it is likely the case that, in the near future, integrative approaches combining several of these features will streamline the candidate gene prioritization process. For example, Carqueijeiro *et al*. [18] combined gene expression and physical clustering analysis to identify an acetyl transferase involved in the biosynthesis of root alkaloids in *C. roseus*. Further automated integrative approaches will undoubtedly facilitate the identification of missing enzymes, as knowledge on plant biosynthetic machineries is rapidly progressing. Because many biosynthetic genes are now fully characterized, this knowledge may indeed drive supervised machine learning (SML)-based approaches to predict gene function [19]. A training set representing 80% of a transcriptome labeled as following: characterized genes or orthologs as 'alkaloid-related' and the remaining ones as 'nonalkaloid'. SML tries to construct the best model from input data to correctly classify genes with the correct label and predicts possible labels for the remaining 20% of genes. Although promising, fine optimization is required to correctly deploy SML approaches for missing gene identification.

It is important to note that although natural product biosynthesis in plants displays a certain form of homogeneity (coexpression, physical clustering, or sequence homology), some steps are catalyzed by enzymes encoded by genes without obvious distinguishing features, making them difficult to detect through these genomic approaches. In this case, many candidate genes identified *in silico* must be tested before the desired catalytic activity is discovered. High-throughput techniques to clone candidate genes or more conventional procedures such as native protein purification through cell fractionation must thus be considered as part of the elucidation process. Based on this whole set of approaches, completion of biosynthetic pathways will accelerate in the coming years and, in turn, will facilitate the development of microbial cell factories synthetizing plant-derived drugs [20,21]. This remains an essential prerequisite to be able to propose, in the short term, cheaper and high-throughput alternatives of production of these valuable pharmaceutical compounds.

Acknowledgments

We acknowledge funding from the EU Horizon 2020 research and innovation program (MIAMi project-Grant agreement N°814645), ARD2020 Biopharmaceutical program of the Région Centre Val de Loire (BioPROPHARM and CatharSIS projects), La ligue Contre le Cancer (Yeast4LiFE), and from le Studium (Consortium fellowship) and ERC 788301. The authors want to apologize for other excellent studies which have not been cited here due to length restriction.

[1]Biomolécules et Biotechnologies Végétales, BBV, EA2106, Université de Tours, Tours, France
[2]Groupe d'Etude des Interactions Hôte-Pathogène, GEIHP, EA3142, Univ Angers, SFR 4208 ICAT, Angers, France

[3]Department of Natural Product Biosynthesis, Max Planck Institute for Chemical Ecology, Jena, Germany

*Correspondence:
oconnor@ice.mpg.de (S.E. O'Connor) and
vincent.courdavault@univ-tours.fr (V. Courdavault).

https://doi.org/10.1016/j.tips.2019.12.006

References

1. Newman, D.J. and Cragg, G.M. (2016) Natural products as sources of new drugs from 1981 to 2014. *J. Nat. Prod.* 79, 629–661
2. Noda-Garcia, L. *et al.* (2018) Metabolite-enzyme coevolution: from single enzymes to metabolic pathways and networks. *Annu. Rev. Biochem.* 87, 187–216
3. Dastmalchi, M. *et al.* (2017) Family portraits: the enzymes behind benzylisoquinoline alkaloid diversity. *Phytochem. Rev.* 17, 1–29
4. Wei, Y. *et al.* (2018) Recent developments in the application of P450 based biocatalysts. *Curr. Opin. Chem. Biol.* 43, 1–7
5. Fox Ramos, A.E. *et al.* (2019) Natural products targeting strategies involving molecular networking: different manners, one goal. *Nat. Prod. Rep.* 36, 960–980
6. Salim, V. *et al.* (2018) *Camptotheca acuminata* 10-hydroxycamptothecin O-methyltransferase: an alkaloid biosynthetic enzyme co-opted from flavonoid metabolism. *Plant J.* 95, 112–125
7. Farrow, S.C. *et al.* (2018) Cytochrome P450 and O-methyltransferase catalyze the final steps in the biosynthesis of the anti-addictive alkaloid ibogaine from *Tabernanthe iboga*. *J. Biol. Chem.* 293, 13821–13833
8. Edge, A. *et al.* (2018) A tabersonine 3-reductase *Catharanthus roseus* mutant accumulates vindoline pathway intermediates. *Planta* 247, 155–169
9. Qu, Y. *et al.* (2019) Completion of the canonical pathway for assembly of anticancer drugs vincristine/vinblastine in *Catharanthus roseus*. *Plant J.* 97, 257–266
10. Chen *et al.* (2018) A pathogenesis-related 10 protein catalyzes the final step in thebaine biosynthesis. *Nat. Chem. Biol.* 14, 738–743
11. Zhao, D. *et al.* (2017) De novo genome assembly of *Camptotheca acuminata*, a natural source of the anti-cancer compound camptothecin. *Gigascience* 6, 1–7
12. Liesecke, F. *et al.* (2018) Ranking genome-wide correlation measurements improves microarray and RNA-seq based global and targeted co-expression networks. *Sci. Rep.* 8, 10885
13. Tatsis, E.C. *et al.* (2017) A three enzyme system to generate the Strychnos alkaloid scaffold from a central biosynthetic intermediate. *Nat. Commun.* 8, 316
14. Lau, W. and Sattely, E.S. (2015) Six enzymes from mayapple that complete the biosynthetic pathway to the etoposide aglycone. *Science* 349, 1224–1228
15. Caputi, L. *et al.* (2018) Missing enzymes in the biosynthesis of the anticancer drug vinblastine in Madagascar periwinkle. *Science* 360, 1235–1239
16. Dang, T.-T.T. *et al.* (2018) Sarpagan bridge enzyme has substrate-controlled cyclization and aromatization modes. *Nat. Chem. Biol.* 14, 760–763
17. Wijekoon, C.P. and Facchini, P.J. (2012) Systematic knockdown of morphine pathway enzymes in opium poppy using virus-induced gene silencing. *Plant J.* 69, 1052–1063
18. Carqueijeiro, I. *et al.* (2018) A BAHD acyltransferase catalyzing 19-O-acetylation of tabersonine derivatives in roots of *Catharanthus roseus* enables combinatorial synthesis of monoterpene indole alkaloids. *Plant J.* 94, 469–484
19. Kim, G.B. *et al.* (2019) Machine learning applications in systems metabolic engineering. *Curr. Opin. Biotechnol.* 64, 1–9
20. Luo, X. *et al.* (2019) Complete biosynthesis of cannabinoids and their unnatural analogues in yeast. *Nature* 567, 123–126
21. Li, Y. *et al.* (2018) Complete biosynthesis of noscapine and halogenated alkaloids in yeast. *Proc. Natl. Acad. Sci. U. S. A.* 115, E3922–E3931