Journal of Plant Biology, March 2002, 45(1) : 1-6 **REVIEW**

Production of Useful Secondary Metabolites in Plants: Functional Genomics Approaches

Jang Ryol Liu[1,2]*, Dong-Woog Choi[2], Hwa-Jee Chung[2], and Sung-Sick Woo[3]

[1]Plant Cell Biotechnology Laboratory, Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejon 305-333, Korea

[2]Laboratory of Functional Genomics for Plant Secondary Metabolism (National Research Laboratory), Eugentech Inc., Daejon 305-333, Korea

[3]UniGen Inc., Daeyang Koreana Bldg, 182-2, Bangyi 1-Dong, Songpa-Gu, Seoul 138-834, Korea

The paradigm of biological research has been changed by recent developments in genomics, high-throughput biology, and bioinformatics. Conventional biology often was based on empirical, labor-intensive, and time-consuming methods. In the new paradigm, biological research e is driven by a holistic approach on the basis of rational, automatic, and high-throughput methods. New functional compounds can be discovered by using high-throughput screening systems. Secondary metabolite pathways and the genes involved in those pathways are then determined by studying functional genomics in conjunction with the data-mining tools of bioinformatics. In addition, these advances in metabolic engineering enable researchers to confer new secondary metabolic pathways to crops by transferring three to five, or more, heterologous genes taken from various other species. In the future, engineering for the production of useful compounds will be designed by a set of software tools that allows the user to specify a cell's genes, proteins, and other molecules, as well as their individual interactions.

Keywords: Bioinformatics, high-throughput biology, secondary metabolism

Plants produce numerous secondary metabolites that have historically been used as pharmaceuticals, fragrances, flavor compounds, dyes, and agrochemicals. Even today, these metabolites are a major source of new drugs. Because their profiles vary among species, extracts from both known and newly discovered plants are subject to screening for use as new medications. Secondary metabolites are usually produced in-vivo only at low concentrations, so large-scale production systems have been developed for use in plant cell culture. Despite the enormous commercial efforts covering many decades, only a few compounds have successfully been produced for less cost than those required for production by direct plant extraction or through chemical synthesis.

As an alternative, the metabolic pathways of useful secondary metabolites have been modified to improve their productivity via genetic transformation. Although this engineering approach seems promising, it has been limited by our current level of understanding of secondary metabolism at the molecular level. This new era of expanding opportunities is having an increased impact on many fields of science and technology. Conventional biology has been changed by recent developments in genomics and bioinformatics, and has shifted from a science often based on empirical, labor-intensive, and time-consuming procedures. In the new paradigm, biology is driven by a holistic approach on the basis of rational, automatic, and high-throughput methods. In this review, we assess the various opportunities for studies of plant secondary metabolism, based on this new paradigm.

*Corresponding author; fax +82-42-860-4608
e-mail jrliu@mail.kribb.re.kr

HIGH-THROUGHPUT SCREENING

Conventional, labor-intensive lead screening has identified only a few hundred useful compounds from plant extracts. Now a new integrated technology, high-throughput screening (HTS), has drastically changed the way this screening is done. HTS uses cell lines, DNA microarrays, and interpretation of data by bioinformatics. It incorporates robotic handling of small amounts of materials, multi-channeled, high-sensitivity chemical analysis of plant extracts, and rapid assay of leads for their therapeutic potentials.

A good example of an HTS system for plant sec-

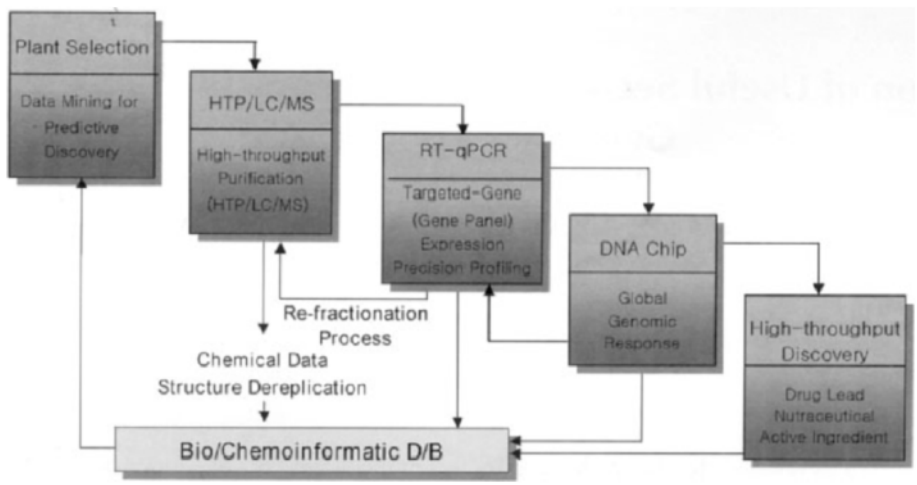

Figure 1. Schematic illustration of the PhytoLogix™ Discovery Process.

ondary metabolites is the PhytoLogix™ Discovery Process (Fig. 1), initially developed by UPI (Denver, CO, USA), and licensed to UniGen, a Korean venture capital company. This new product represents the first integrated process for discovering drug leads, nutraceuticals, and new therapeutic applications from natural products, especially plants. The process consists of high-throughput systems for: 1) purification of chemical compounds from plant extracts, including automatic fractionation, LC/MS analysis, and data collection for chemical fingerprinting and mass profiling of target compounds; 2) precision profiling of individual compounds or fractions through qPCR-based detection of relative mRNA expression with proprietary targeted-gene panels; and 3) differential gene-expression profiling that uses DNA-microarray screening for global activity patterns and further improvement of those proprietary gene panels in a hierarchical fashion. The quantitative-profiling results are analyzed and compared to known drugs to establish a mechanism of action, a safety/toxicity profile, metabolism, and potency for evaluation as a drug lead for commercialization. Information from this process is merged into a sophisticated relational database that combines public and proprietary information from the fields of traditional medicine, genomic analysis, human clinical testing, and chemical analysis of ethno-medicinal plants found worldwide. This database is also the information source for predictive discoveries from ethno-medicinal plants.

UniGen is currently focused on identifying anti-arteriosclerosis ingredients, using expression-profiling of a proprietary gene panel specific to arteriosclerosis in vascular endothelial and macrophage cell lines. The emphasis at UPI is on discoveries from medicinal plants of the eastern and western hemispheres. Using the high-throughput purification system (HTP/LC/MS) at the beginning of the PhytoLogixTM Discovery Process, a total of 50 plant extracts can be fractioned into 96-well deep plates/module in one day, while gathering chemical information from each well at several collection points. In this way, the company will be able to collect data and build a natural-compound library for 40,000 medicinal plants from around the world by 2005.

EST APPROACH

The partial sequencing of anonymous cDNA clones, known as expressed sequence tags (ESTs), has been widely used as a rapid and cost-effective means of obtaining information about gene expression and the coding capacity of plants. ESTs expressed in specialized tissues and organs have been analyzed to assist in identifying new genes involved in specialized pathways. Lange et al. (2000) directed their functional genomics approach toward characterization of the genes involved in essential oil (monoterpene) formation in peppermint (*Mentha* x *piperita*). They initiated their study with the construction of an oil-gland secretory-cell cDNA library. These secretory cells are responsible for essential oil biosynthesis and can be isolated in large quantities from the leaves. Isolated cells are capable of de-novo biosynthesis of monoterpenes from primary carbohydrate precursors. The EST sequences from this cDNA library (1,300 acquisitions) have led to bioinformatic processing of the data and putative identification of the candidate genes

involved in essential oil biosynthesis. Subsequent secretions have enabled functional characterization of the corresponding recombinant proteins that are overexpressed in *Escherichia coli*. This research represents an important step toward development of a metabolic map for oil glands, and provides a valuable resource for defining the molecular targets for genetic engineering of essential-oil formation.

BIOCHEMICAL GENOMICS APPROACH

When a gene produces an extremely small amount of protein, its purification can be difficult, time-consuming, and expensive, but that step is often a prerequisite for cloning the gene and assessing its function. Martzen et al. (1999) have designed a new in-vitro biochemical screening method for determining activity in cDNA or an open reading frame (ORF; Fig. 2). An array of 6,144 individual yeast strains has been generated, each expressing a fusion protein between glutathionine S-transferase (GST) and an individual ORF from yeast under the control of a copper-inducible promoter. This array of yeast strains was cultured in defined pools (64 pools of 96 GST-ORF fusions). Expression of the chimeric proteins was induced by adding copper sulfate to the cells, and extracts were prepared from each pool.

Because the GST-tag was present, the expressed chimeric proteins were readily purified, to a high degree, by glutathione-agarose affinity chromatography. Biochemical assays were then performed, using these purified proteins to find an ORF of interest. When a positive result from a particular pool was obtained, it was necessary to return to the 96-well plate to create the pool and re-assay the rows and columns of that plate. This secondary assay procedure was then used to identify the intersection point between row and a column in the plate, thus indicating the precise ORF. With this approach, the efficiency and technical

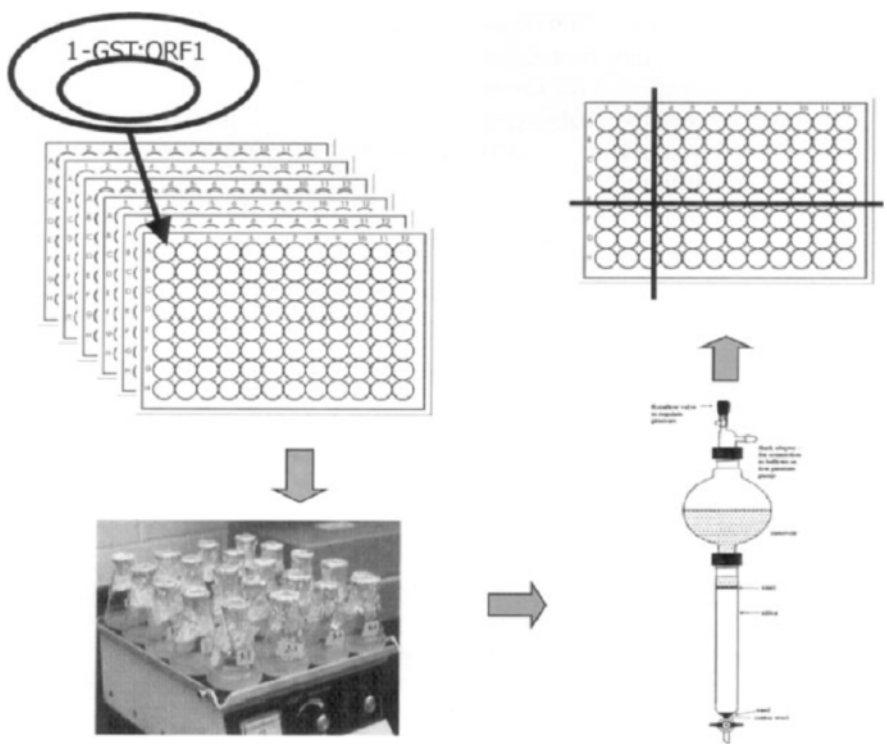

Figure 2. Schematic illustration of the biochemical genomics approach. A: Each yeast ORF is fused to GST under control of the CUP1 promoter, and the constructed vector is introduced into yeast. The resulting set of 6144 yeast strains is placed into 64 microtiter plates of 96 wells each. B: To obtain the set of GST-ORF proteins, the 96 strains on each plate are pooled into 64 se ts before being cultured. C: Each cultured set is subjected to lysis, then the GST-ORFs are purified by glutathione-agarose affinit y chromatography. Protein pools are assayed for biochemical activity to determine the microtiter plate containing the strain of interest. D: Entire procedure is repeated with the plate containing 64 strains, for which the pooled products exhibit positive activity. Again, the 64 strains are pooled by row and column of the plate to be cultured and lysed. Then, the strain whose product was detected in a pooled row and column with positive activity simultaneously is identified.

advantages of working with pools and 96-well plates provides for the identification of ORF-associated activity in less than two weeks.

METABOLOMICS APPROACH

The most reliable means for determining precise gene function is by comparing a wild-type organism that naturally expresses that gene with mutants that express the gene at either low or high levels. These morphological comparisons frequently are followed by comparisons both at the transcriptional level using a DNA microarray, and at the protein level via two-dimensional gel electrophoresis. Such an analysis can also be made at the metabolite level by profiling small molecules, as Fiehn et al. (2000) have demonstrated. Their sophisticated analytical technique was developed to identify and monitor a large number of compounds in extracts from *Arabidopsis thaliana*. Using a relatively simple extraction method coupled with two-step derivatization and analysis by gas chromatography/mass spectrometry (GC/MS), they have been able to yield the relative concentrations of hundreds of compounds in plant extracts. Selected *Arabidopsis* genotypes confer metabolic profiles distinct enough to conduct such a comparative analysis.

Most of the small molecules that Fiehn et al. (2000) used to obtain their metabolic profiles were amino acids and sugars. However, secondary metabolites in plant extracts can also be detected and can provide distinct profiles. Therefore, the large volume of profiles collected from mutants in a biased analysis (e.g., terpenoids only), coupled with data-mining tools, will enable us to determine the secondary-metabolite pathways as well as the functions of the genes involved there.

DNA microarray data are often subject to clustering for data mining. This technique allows a set of genes to be divided into related subsets based on a distance metric. In the future, metabolic-profiling data will also be subject to this cluster-analysis. However, that method cannot explain dependency and causality among genes. Using Bayesian Networks, we may be able to overcome these limitations and elucidate the pathways of secondary metabolites from a large volume of data. Change et al. (2001) have demonstrated this in their analysis of gene-drug dependency.

METABOLIC ENGINEERING

Metabolic engineering is generally defined as the redirection of one or more enzymatic reactions to produce new compounds within an organism, to improve the production of existing compounds, or to mediate the degradation of those compounds (DellaPenna, 2001). One of the best examples of this is from Yun et al. (1992), who genetically modified *Atropa belladonna* (Solanaceae). They introduced the hydroxylase gene, which catalyses formation of scopolamine from hyoscyamine, into *A. belladonna* via *Agrobacterium*-mediated transformation. A transgenic plant was then selected that strongly expressed the transgene. Because this plant and its selfed progeny contained high concentrations of both atropine and scopolamine, Japanese pharmaceutical companies are now using these plants as their major source of those compounds.

Metabolic engineering is based on plant genetic transformation and molecular dissection of metabolic pathways. Transgenic rice, which contains a high level of carotene (provitamin A) in the endosperm, has been generated by introducing three genes (two from plants and one from bacteria) that encode for the carotenoid biosynthetic enzymes (Ye et al., 2000). This so-called "golden rice" contains 10% of the recommended daily carotene allowance in one's average daily intake. Conventional rice is a naturally poor source of carotene, and vitamin A deficiency is a serious health issue in developing countries where rice is the staple food. Advances in genomics and related studies for secondary metabolism continue to promote more sophisticated metabolic engineering of plants.

AN INTEGRATED APPROACH: A CASE OF FUNCTIONAL GENOMICS FOR SECONDARY METABOLISM USING GINSENG HAIRY ROOT CULTURES

Korean ginseng root contains manyl desirable compounds, including the glycoside panaquilon (a saponin panaxin), oils, vitamins B1 and B2, alkaloids, and polysaccharides. These ingredients have various pharmacological effects on the human body, as well as antioxidant and anticancer activities. Therefore, ginseng can be a source of useful and effective drugs.

To clone a large number of genes for secondary metabolism and to map the metabolic pathways of useful compounds in this species, scientists at Eugentech Inc., a Korean venture capital company, have recently launched a project for functional genomics of its secondary metabolism (Fig. 3). This research is modeled after mutation-based approaches in order to study metabolism in bacterial systems. Hairy root cul-

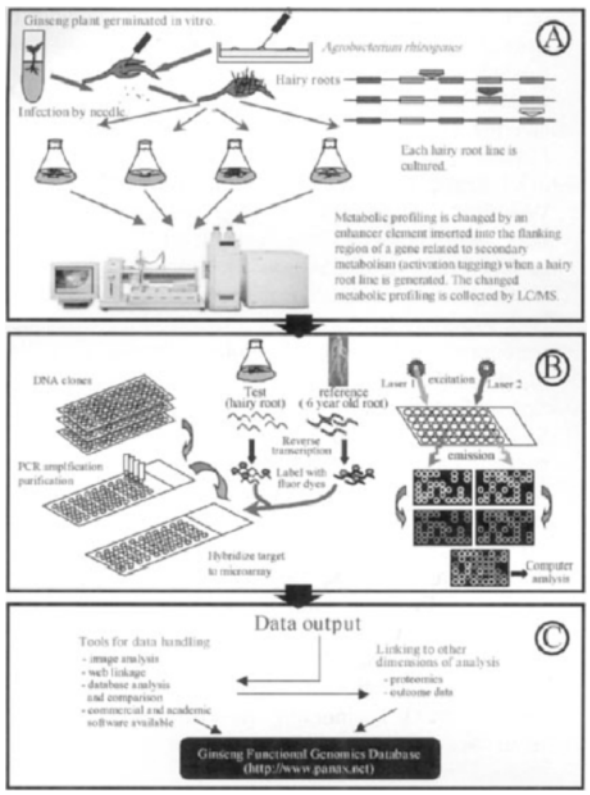

Figure 3. Schematic illustration of functional genomics for secondary metabolism of ginseng hairy root cultures. Metabolic profiling data from LC/mass analysis of each ginseng hairy root line are combined with gene-expression profiling data from DNA microarray analysis to elucidate genes involved in secondary metabolism.

tures of ginseng are used for transgenic mutant analysis because most of the useful secondary metabolites are produced in the root. Numerous hairy root cultures are easily generated from leaf tissues through inoculation with *Agrobacterium rhizogenes*, which harbors the Ti-plasmid with an activation-tagging sequence. At least two approaches are combined in this project. First, LC/MS is used to characterize numerous hairy root lines on the basis of metabolic profiling. Second, each hairy root line that is so metabolically characterized is then subject to DNA microarrays that consist of ESTs specific to secondary metabolism. The large volume of comprehensive metabolic-profiling and DNA-microarray data that is generated in this process will be subjected to dependency analyses of gene-gene expression, gene expression-secondary metabolite activity, and secondary metabolite-secondary metabolite activity, using Bayesian Networks to determine the pathways and genes involved in secondary metabolism.

CONCLUSIONS AND FUTURE PERSPECTIVES

For the last few decades, scientists in the field of plant secondary metabolism have worked on: 1) discovering new functional compounds; 2) identifying the metabolic pathways of secondary metabolites; 3) identifying genes for the enzymes that catalyze those pathways; 4) producing useful secondary metabolites in cell-culture systems; 5) metabolically engineering secondary metabolism at the whole-plant level; and 6) scaling-up those cell-culture systems and purifying useful secondary metabolites for commercialization. All these efforts have involved empirical, labor-intensive, time-consuming conventional methods.

Since a holistic approach to plant secondary metabolism was introduced, based on genomics, high-throughput biology, and bioinformatics, the paradigm for such research has been drastically changed. New functional compounds can now be discovered via HTS. Likewise, the pathways for secondary metabolites and the genes involved in those pathways can be determined through functional genomic approaches in conjunction with data-mining tools. In addition, plants can be metabolically engineered with three to five, or more, heterologous genes for large-scale production of useful secondary metabolites. For example, a few heterologous genes under the control of a promoter as an operon can be introduced into a plastid. The gene products are then be safely compartmentalized from the degrading enzymes found in the cytosol.

E-CELL is a model-building kit (Tomita et al., 1999). This set of software tools allows a user to specify a cell's genes, proteins, and other molecules; describe their individual interactions; and then compute how they work together as a system. This kit, pioneered by Masaru Tomita, a professor of bioinformatics at Keio University in Fujisawa, Japan, should ultimately allow investigators to conduct experiments "*in silico*". It offers an inexpensive, rapid means for screening candidate drugs, studying the effects of mutations or toxins, or simply probing the networks that govern cell behavior (Normile, 1999). Tomita has just completed a model of human erythrocytes, and is building other models of human mitochondria, signal transduction for chemotaxis in the bacterium *E. coli,* and gene-expression networks in this bacterium's lactose operon (Butler, 1999). Although plant cells are genetically much more complicated than are bacterial cells (rice genome: 430 megabases vs. *E. coli* genome: 4.6

megabases), it does seem possible to soon play with a "tamagotchi" plant in the near future. E-CELL may represent highly sophisticated metabolic engineering for future production of useful crop compounds.

ACKNOWLEDGEMENTS

This work was supported by a grant (M10104000234-01J000-10710) to JRL from the National Research Laboratory Program; a grant (PF003101-04) to DWC from the Plant Diversity Research Center of the 21st Century Frontier Research Program; a grant to JRL from the Korea Science and Engineering Foundation through the Plant Metabolism Research Center (Kyung Hee University); and a grant (No. CGM0400111) to JRL from the Crop Functional Genomics Center of the 21st Century Frontier Research Program, funded by the Ministry of Science and Technology, Republic of Korea.

Received March 9, 2002; accepted March 10, 2002.

LITERATURE CITED

Butler D (1999) Computing 2010: From black holes to biology. Nature 402: C67-C70

Chang JH, Hwang KB, Zhang BT (2001) Analysis of gene expression profiles and drug activity patterns for the molecular pharmacology of cancer. In proceedings of Critical Assessment of Techniques for Microarray Data Analysis (CAMDA'01), October 15-16, 2001, Duke University, Durham, NC

DellaPenna D (2001) Plant metabolic engineering. Plant Physiol 125: 160-163

Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RT, Willmitzer L (2000) Metabolite profiling for plant functional genomics. Nat Biotechnol 18: 1157-1161

Lange BM, Wildung MR, Stauber EJ, Sanchez C, Pouchnik D, Croteau R (2000) Probing essential oil biosynthesis and secretion by functional evaluation of expressed sequence tags from mint glandular trichomes. Proc Natl Acad Sci USA 97: 2934-2939

Martzen MR, McCraith SM, Spinelli SL, Torres FM, Fields S, Grayhack EJ, Phizicky EM (1999) A biochemical genomics approach for identifying genes by the activity of their products. Science 286: 1153-1155

Normile D (1999) Complex systems: Building working cells '*in silico*'. Science 284: 80-81

Tomita M, Hashimoto K, Takahashi K, Shimizu T, Matsuzaki Y, Miyoshi F, Saito K, Tanida S, Yugi K, Venter JC, Hutchison C (1999) E-CELL: Software environment for whole cell simulation. Bioinformatics 15: 72-84

Ye X, Al-Babili S, Kloti A, Zhang J, Lucca P, Beyer P, Potrykus I (2000) Engineering the Provitamin A (β-carotene) biosynthetic pathway into (carotenoid-free) rice endosperm. Science 287: 303-305

Yun DJ, Hashimoto T, Yamada Y (1992) Metabolic engineering of medicinal plants: Transgenic *Atropa belladonna* with improved alkaloid composition. Proc Nat Acad Sci USA 89: 11799-11803