Plant Metabolic Engineering for High Value Products

Modelling and Databases for the Enhancement of Terpenes in Plants

WG1 Workshop February 16th - 18th 2012 Wageningen, The Netherlands

About COST

COST (European Cooperation in Science and Technology) is one of the longest-running European instruments supporting cooperation among scientists and researchers across Europe. COST key features are: building capacity by connecting high-quality scientific communities throughout Europe and worldwide; providing networking opportunities for early career investigators; increasing the impact of research on policy makers, regulatory bodies and national decision makers as well as the private sector. COST is a building block of the European Research Area, instrumental for successful innovation strategies and global cooperation. More information about COST can be found at http://www.plantengine.eu/

About COST Action FA1006 - Plant Metabolic Engineering for High Value Products (PlantEngine)

Plants produce a vast variety of chemicals for various purposes. Those "Plant Natural Products" -or PNPs- are often of special interest since they can serve as fragrances, fine chemicals or as pharmaceuticals. Chemical synthesis of the usually very complex structures is theoretical feasible but often not applicable for large-scale production at reasonable prices. Hence, plants are still in many cases the sole source for PNPs. Given that some PNPs are only produced in minute amounts or that the host plant is a slow growing or even an endangered species, there is a clear need to find ways to engineer plants to optimize the production of the "high value compounds".

Although a tremendous amount of knowledge has been gained during the last decades about the biosynthetic capacity of plants and the pathways leading to the formation of PNPs, there are still many white spots on the maps describing the biosynthetic sequences. Moreover, the regulation of those pathways in the context of environmental and developmental changes is only poorly understood.

This COST Action will support and enhance a Pan-European network which will amalgamate resources, define target pathways and prioritize compounds, disseminate novel technologies and applications, set standards for computational support, and develop synthetic approaches in plant metabolic engineering. Due to its multidisciplinary approach this Action will initiate a European network of experienced as well as early stage researchers which will serve as a base for future research collaborations. Outcomes will help guiding researchers in the design of plants as production host and provide building blocks for pathway engineering. The dialog established within the research community will involve key players in industry, stakeholders, and policy makers guaranteeing the highest momentum for the European research sector as well as the public.

Program

Thursday	**16.2.**	
17:00-19:00	Welcome and registration	
19:00-21:00	**Dinner**	
21:00-21:10	Introduction by host	S. van der Krol
21:10-22:00	**From sequence to pathways** (Plenary lecture) Reconstruction of metabolic pathways in medicinal plants	R. Mahadevan/ Chris Gowen
Friday	**17.2.**	
9:00-9:15	Introduction to the aims and goals of PlantEngine and WG1	H. Warzecha
9:15-9:45	round table introduction	
	Modelling pathways from metabolomics information	
9:45-10:00	METAPRO	P.Fraser
1000-1030	metabolic pathway flux modelling	Bernd Lange
1030-1045	metabolic pathway modelling	Kay Hamacher:
1045-1100	Competition in the sesquiterpene biosynthesis pathway	S. van der Krol
1100-1115	**Coffee break**	
1115-1130	Modelling volatile isoprenoid emission in relation to stress	Melanie Senning
1130-1200	The complexity of integrating transcriptomics, metabolomics and proteomics	Ronnie Joosen
1200-1215	Transcriptional correlation for pathway reconstruction	Staffan Persson
1215-1245	Intermediate conclusion	
1245-1400	**Lunch**	
	Modelling protein structure for activity specificity	
1400-1430	Modelling P450 enzyme activity	Jean Francois Ginglinger
1430-1500	Molecular and structural reasons of metabolic diversity mediated by prenyl converting enzymes	Ludger Wessjohann, Wolfgang Brandt
1500-1515	**Coffee break**	
	Data storage and sharing	
1515-1530	Chemical and Plant Ontology for Plant Natural Products Data Mining	Lutz Weber
1530-1545	The Smartcell database	Virkki Arho
1545-1615	AtIPD: A Curated Database of Arabidopsis Isoprenoid Pathway Models and Genes for Isoprenoid Network Analysis	Eva Vranovà-Milcakova:

1615-1800	**Define questions and group discussion**
1800-1900	Free – recreation and individual discussions
1900-2030	**Diner**
2030-2200	Report and summary of discussion groups, general discussion. Prioritisation and Realisation opportunities-final conclusions

Saturday **18.2.**
Departure

9:00-11:00	Core group: report and layout opinion paper

Participants:

First Name	Last Name	Country	Email
Wang	Bo	Netherlands	bo2.wang@wur.nl
Harro	Bouwmeester	Netherlands	harro.bouwmeester@wur.nl
Wolfgang	Brandt	Germany	Wolfgang.Brandt@ipb-halle.de
Aristotelis	Chatziioannou	Greece	achatzi@eie.gr
Henrik	Cordes	Germany	henrik.cordes@googlemail.com
Paul	Fraser	United Kingdom	P.Fraser@rhul.ac.uk
Jean-Francois	Ginglinger	France	jfginglinger@unistra.fr
Chris	Gowen	Canada	chris.gowen@utoronto.ca
Kay	Hamacher	Germany	hamacher@bio.tu-darmstadt.de
Björn	Hamberger	Denmark	bjoernh@life.ku.dk
Ronny	Joosen	Netherlands	Ronny.joosen@wur.nl
Bernd Markus	Lange	United States	lange-m@wsu.edu
Dong	Lemeng	Netherlands	lemeng.dong@wur.nl
Staffan	Persson	Germany	persson@mpimp-golm.mpg.de
Michael	Phillips	Spain	michael.phillips@cragenomica.es
Melanie	Senning	Switzerland	msenning@ethz.ch
Sander	Van Der Krol	Netherlands	sander.vanderkrol@wur.nl
Aalt Jan	Van Dijk	Netherlands	aaltjan.vandijk@wur.nl
Arho	Virkki	Finland	Arho.virkki@vtt.fi
Eva	Vranova	Switzerland	evranova@ethz.ch
Heribert	Warzecha	Germany	warzecha@bio.tu-darmstadt.de
Lutz	Weber	Germany	lutz.weber@ontochem.com
Ludger	Wessjohann	Germany	Ludger.Wessjohann@ipb-halle.de

Program:

Thursday, 16.2.

The meeting was opened by the host and local organizer, Sander van der Krol. He expressed his welcome to the 23 participants from very diverse fields.

To introduce collaborative activities in the field of PNP outside of Europe, **Chris Gowan** presented the Canadian consortium PhytoMetaSyn (www.phytometasyn.ca) and his work on *Metabolic System Engineering*. This consortium of 13 groups with a budget of 14,6 mio $ has the objectives to identify biosynthetic steps in PNP biosynthesis and to develop commercial scale production systems. One part of this effort, the development of platform strains for the efficient production of intermediates in PNP biosynthesis was discussed in detail.

Friday, 17.2.

After an introduction to COST and the aims and goals of COST Action FA1006, PlantEngine by the MC Chair, **Heribert Warzecha** (D), all participants introduced themself as well as their main research foci. Here it became already clear that participant covered the whole scale from wet lab to dry lab, and also covering topics beyond isoprenoids.

Paul Fraser (UK), Vice Chair of this COST Action and coordinator of METAPRO introduced the research of the consortium on carotenoid biosynthesis and metabolic engineering. Special focus was put on to the fact that metabolic engineering frequently causes unintended negative effects and that very little is known about the causes. Also, beside the increase of compound content the storage is a very important issue and the engineering of plastid number and size is one focus of this consortium. Paul discussed how modelling of the metabolic network identified several 'hubs' which were thought to be important targets for manipulation. But it turned out that hubs are actually very difficult to manipulate because of the interaction with so many different factors, and manipulation of targets one step away from the hub were far more effective.

Mark Lange (USA) was presenting his work on monoterpenoid engineering in peppermint, again an issue in which storage of the engineered compounds become important. In this case glandular trichomes are the target of engineering. He pointed out that also with a pathway as well understood and characterized as the one leading to

monoterpenoids, extensive measurements are mandatory to successfully attempt engineering. He showed that production of essential oils was not easily modelled unless development of trichomes on leaves was taken into account. The capacity to store essential oil levels per trichome were more or less fixed and thus oil yield were a direct function of trichome number and development.

Kay Hamacher (D) introduced the computational biology approaches to use data gathered in the lab to model and analyse biosynthetic networks exemplified by fluctuation analysis in glycolysis. Also, during the talk the requirements for a database as well as the format of data were discussed.

Sander van der Krol (NL) again switched back to the wet lab approaches to transiently assemble pathways in heterologous hosts step by step to analyse the detailed function of biosynthetic enzymes and provide quantitave data for model input. Differences in protein structure of P450s in the artemisinin pathway are currently being related to difference in product range of these two proteins. One of the difficulties in modeling pathways in plants is the different subcellular location of biosynthesis enzymes and the unknown exchange of metabolites between compartments. The wet experiments in pathway reconstruction targeted to different subcellular compartments will be compared to model outputs and may give clues about the substrate exchange rate between compartments. The relation of different kinds of stress to the production and emission of volatile isoprenoids were discussed by **Melanie Senning** (CH) and **Ronny Joosens** (NL) described the QTL mapping for Arabidopsis seed formation and identification of master switches for metabolite formation. By combining seed germination phenotypes with eQTL, metabolomics and proteomics data different approaches to network analysis and visualisation was used to browse through such large data sets and extract biological relevant interacting components. **Staffan Persson** (D) described the approach to identify non-known components in cellulose and cell wall biosynthesis by identifying reduced networks within large networks. The correlation networks can be dissected by defining 'island' of interaction, depending on connections within and connections between the group and the rest of the network. An expansion to plants like Medicago will eventually enable the use of this method for PNP pathway investigation. The plant coexpression network browser is available (http://aranet.mpimp-golm.mpg.de)

Jean Francois Ginglinger (F) presented details of the global approach to elucidate Cytochrome P450 function and the CYPedia tool which is available at http://www-ibmp.u-strasbg.fr/~CYPEdia/ . The P450s make pathway modelling difficult but at the same time also very necessary; a combination of promiscuous substrate specificity and liberal product output with even single substrate input, makes it difficult to predict P450 enzyme function from sequence information.

Also with CYP450 enzymes in focus, **Björn Hamberger** (DK) described the efforts to unravel the forskolin biosynthesis, favouring the term metabolic grid over metabolic pathway since in plant metabolism, numerous side reactions occur and a diverse set of products is formed. Moving more into biochemistry and protein modelling, **Ludger Wessjohann** (D) discussed the thermodynamics of geranyl cation formation. Here he pointed out that special emphasize should be given to correct biochemical characterization of enzymes as well as to proper structural data for chemicals.

Lutz Weber (D) as a company representative described the potential of text mining tools based on onthologies and introduced a program to retrieve data from various sources.

Referring again to the computational part of this meeting **Arho Virkki** (FI) described how the SmartCell consortium tries to harmonize data acquisition and storage to obtain a unique database for the project. Moving back to the interface between wet and dry lab, **Eva Vranova** (CH) introduced a curated database for isoprenoid pathways in Arabidopsis (www.atipd.ethz.ch)

In the first discussion after the presentations it became clear that everybody sees the benefit of an integrated solution but a clear consensus about the "how" and "what" will not be achieved easily.

Problems with metabolomics data:

A clear difference can be seen between the datasets which could be integrated into a database: while genomics as well as proteomic data are more defined, metabolomics data become extremely complex (because of the complexity of metabolites and because of the different techniques required to measure different types of metabolites). Beside the variation in instrumentation-dependant data acquisition and storage modes, also the size when raw data are to be stored might cause problems. Metabolomics data is often stored in processed form. For instance, after alignment of the chromatogram of different

samples the data are reduced to a list of masses with intensity signal in the different samples. Because the alignment may introduce false positive results (alignment of peaks that do not belong to the same compound) it is always necessary to go back to the original chromatograms, or to verify results in independent experiments. Potentially, alignment procedures could improve over the years, for which it than would be useful to reprocess old data to extract more or more reliable results. In this context it was suggested that maybe the common bulk primary metbolites could be used as an internal standard for alignment, which actually eventually also could enable alignment across different platforms.

Should there be rules how data needs to be structured

There is already a MIRIAM protocol for metabolomics data, similar to the minimal requirements for microarray data. Often these protocols are viewed as too restrictive when applied to the actual raw date. However, for a good sample description, a well structured metadata is still essential for intelligent use of stored data.

Database curration and evolution:

It was clear to everyone that current databases need improvement and updating. Current databases, although regularly updated, face several problems. Curation is often lagging, data are inaccurate or selection criteria are not well defined. Such problems were noted for the KEGG database but also for the AraCyc database. Moreover, the new requirements/possibilities are often more efficiently implemented into a new database (e.g. the arabidopsis isoprenoid dedicated DB atipd) instead of updating existing DB's. Maybe we have to accept that each time a DB can only be set up optimally with the current knowledge and that experience shows that in maybe in a few years from now novel features are easier implemented in a newly structured DB than updating an existing DB. Still, in the development of new DBs the existing DBs have a prominent function in providing the basic content. Old and new DBs are subject to evolutionary selection, outdated DB's will automatically fade out of existence but may for long time still provide inaccurate data. New DBs only survive by people recognizing the advantages of the new DB and (as Kay suggested).

Structural information versus biological information:

From the chemical field it was noted that in many DBs the structural information and related spectral supporting data are missing and many DBs even report wrong structures. Although the biology field fully underlines the importance of sound structural information the nature of metabolomics data acquisition does not allow for full spectral support of each individual compound (especially for LCMS). Rather, in the biological field the focus is on gathering biological information on unknown compounds (unknown masses) though correlation analysis in bioassays, tissue specific expression etc. Rational behind this approach is that the investment in full elucidation of composition and structure is only worthwhile when the compound is linked to a clear biological activity. It was noted that in these correlation-type of experiments (e.g. linking the presence of a mass with certain gene expression) could actually help elucidate the compound structure as identification of the gene and encoded enzyme may give clues about the unknown compound itself.

Reliability index

Full integration of 'omics' data already suggests the existence of networks or grids rather than pathways. With the new network analysis tools the sub-network size may be chosen depending of the correlation threshold and reliability of individual connection could be indexed on their correlation strength. Similarily, reliability of compound identification, gene annotation, protein characterisation etc could be indexed. This way a DB could be queried at different levels of reliability, depending on the expertise of the researcher.

Open source natural plant product DB

Several independant initiatives are already in place to collect natural plant compound information (e.g. Solanecea metabolomics DB Cornell University, Tomato metabolite DB PRI, Wageningen, Arabidopsis metabolites, Golm, Massbank, RIKEN Japan, etc). Clearly every one could benefit from cross species validation of natural compounds by combining or exchanging these data. This way, many common natural occurring metabolites do not have to be characterized over and over again at different locations in the world.

Saturday, 18.2.

With the input from the previous day´s talks as well as from the final discussion a smaller group of participants tried to define needs and a putative structure for a PNP database. Also, potential founding sources for such a program were discussed. Several areas and associated constrains were identified:

a) Beside the simple acquisition of data their curation will be crucial. Several examples of useful databases are available but there use is limited due to outdated or inaccurate content (e.g. KEGG).
b) For modelling, different modes were suggested, e.g. co-expression network and ODE networks
c) For the use of metabolomic data, the lack of a robust technology, especially for LC was discussed

In the end, a database of use for all groups would link "real" chemistry data (heavily curated) with metabolomics and genomics/proteomics data. To further define such a structure, questions which could be addressed with such a database are collected:

- For a given reaction step (chemical compound x converted to y): are there other enzymes described (onthology search?) (HW)
- Metabolic profiles under various conditions → connecting spectral data and structures, linking to biological, genetics and biochemistry
- Currently, no database on PNPs is available
- Linking of database to virtual marketplace

Participants discussion points

#relevant data that the database links to could include:

- cellular localization of enzymes and of reactions
- available enzymatic parameters from e.g. Brenda
- enzyme expression levels (preferably both RNA and protein)
- post-translational modifications and/or protein-protein interactions which might influence enzyme activity

#if data is available for a pathway in different species and/or ecotypes it would be important to see the link between those different datasets (e.g. based on orthology) somehow indicating where data originate from and/or how reliable they are is important: e.g. distinction between data directly submitted by users of the database, curated data, data obtained from literature etc.

Participants discussion points

The term modelling appears to be the most widely applied term to a broad range of activities. For the approaches "Plant Pathway Discovery" and the subsequent activity of "Metabolic Engineering" I suggest the following questions and opportunities of high relevance, and areas where databases are already of critical help, but need further development and integration:

(i) Modelling of hypothetical biosynthetic routes, based on evidence derived from the end product (i.e. target molecule), metabolomic analysis (i.e. intermediates and further modified compounds carrying the target backbone) and identification of enzymatic steps supported by transcriptomic or genomic data (i.e. comparative genomics and phylogenetic analysis)

(ii) Deep and comprehensive cross-referenced databases with correlation of metabolomic data, (proteomic data [comment: challenge without genomic or transcriptomic resources]) global sequence data, transcript data (quantitative, across series of tissues/treatments, or cross species comparative) are typically built from scratch for every project. As we slowly begin to explore the plant kingdome at greater depth with more non-model species, integration of previously generated knowledge into the novel databases would allow accelerated pathway discovery and rapid classification of novel candidate genes.

(iii) Metabolic fluxes in engineered and pathways and their endogenous shunt pathways, as well as the endogenous routes toward the precursor molecules need to be effectively modelled to identify and alleviate putative bottlenecks toward the biosynthetic production of novel target compounds in heterologous hosts.

(iv) Novel, engineered combinations of parts of our exponentially growing biosynthetic toolbox allow assembly of new-to-nature pathways. These hold the promise to access to far greater chemical diversity than found in the original plant species, as no selective pressure rests on the pathways other then through the researcher. While so far mostly trial-and-discovery, the rational modelling of the outcome of new combinations using comprehensive databases will accelerate these approaches in Synthetic Biology.

The pathways upstream of the precursor molecules for metabolic engineering have been investigated in a number of trial and error studies aimed at boosting the flux in the most common host organsisms. However, our understanding of the regulation of the flow of metabolites and of the competition of other (vital) endogenous pathways for the central intermediates so far limits effective prediction and modulation of key regulatory steps in less conventional biotechnological host species for terpenoid production such as Physcomitrella patens, algae or cyanobacteria. Fully sequenced genomes in those species, together with existing transcriptomic data could be mapped onto the existing orthologous pathways to model these routes (i.e. modelling of the routes). Identification of shared modules in these biochemically equivalent routes may lead to strategies analogue to those developed for yeast and *E. coli*.

Nomenclature and consistent classification did not raise any concerns, with the few participants discussed. The low-level annotation (class I or II), together with some structural features of the regarding TPS may be sufficient. The P450s are taken care of anyway.

I see a serious bottleneck in the integration and correlation of metabolomic data and sequence data (genomic and transcriptomic). Only a fraction of the researchers aims at closing the 'soft' gap constituted of cross-species comparisons on sequence level [phylogenies], proteomics, hypothetical routes and the discovery of (pharmacologically inactive) intermediates between the two 'hard' layers. Depending on the direction of the approach, these are termed forward (or bottom-up) or reverse (top-down) genetics. Phytochemical approaches on the other hand, further investigate physiological targets through bioassays, and are so far independent lines of research. The common theme

between the two parties is the overlapping interest in solid phytochemistry, with the functional genomics going beyond bioactive metabolites.

Layers:

1 Genomic sequence information and identification of family representation

2 Transcriptomic data distinguishes activity levels across tissues and treatments and helps identifying active players

2b Cross-species comparisons allow identification of genera, or family and species specific genes

3 with 1 and 2 established, proteomic studies yield more direct evidence supporting families and activities

4 Based on 6 and 1-3 hypothetical routes can be developed, which serve as guide for the functional characterisation of the enzymes/genes putatively involved

5 Metabolomics assisted identification of predicted metabolic intermediates

6 Phytochemical information, metabolomics and typical phytochemistry to identify bioactive components in plants

7 Bioassay guided assay of activity towards physiological targets

Directions: 1 to 6 Reverse Genetics; 6 to 1 Forward Genetics;

6 to 7 Phyto-pharmacology

Some more comments:

USER-FRIENDLY CURATION and MODULARITY should be guaranteed by the DATABASE

How to do it:

1) gene network can be taken as a scaffold and filled in gradually with the missing patwhays (genes) from other species (similarly to KEGG).

2) Other type of data should be plugged in (enzymes, metabolites….)

3) If data are taken from other resources which are updated, automatic update function should be implemented and discrepancies between different databases can be highlighted and offered to community as problem to be solved.

3) 1 and 2 can by downloaded from many databases (AraCyc, MetaCyc, KEGG, new Medicinal Plant databases (ETS seqs,, new sequencing data) etc. and community should have option to curate it and add new data.
4) In addition, tools to work with the data should be implemented.

Other keywords: NSF iPlant; Community-driven database vs. Big solution.
User-friendly, modular – both towards curation (community) and usage (data selection and tools to work with them).