Chimia 59 (2005) 362–365
© Schweizerische Chemische Gesellschaft
ISSN 0009–4293

Synergy at the '*Ecole de Pharmacie Genève-Lausanne*': Methodology Developments for the Treatment of Complex Metabolomic Datasets with Data Mining

Aly Thiocone[a], Elia Grata[a,c], Julien Boccard[b], Pierre-Alain Carrupt[b], Serge Rudaz[c], and Jean-Luc Wolfender[a*]

Abstract: With the advances in analytical techniques and data mining, the chances to elucidate a plant metabolome and understand the metabolite variation in response to external stimuli such as stress are gradually becoming feasible in a global manner. This approach represents a very important research challenge because of the structural diversity of the metabolites and the convolute nature of biological matrices. Based on a new collaborative framework emerging from the recent creation of the EPGL, a project aimed at the development of methodology for the treatment of complex metabolomic datasets with data mining has been initiated with the expertise of different EPGL laboratories. In this paper the strategies used for the study of the metabolic variations in a biological model (the effect of mechanical wounding on *Arabidopsis thaliana*) are described. Metabolite profiling based on micro-extraction and LC/MS analysis with various detection methods has been used. Data were parsed in the form of ion maps and treatment of this complex set of MS information was performed with dedicated bioinformatic tools. This approach should enable the evaluation of metabolome variations in a comprehensive manner for a better understanding of complex biological mechanisms and for the detection of novel bioactive molecules.

Keywords: *Arabidopsis thaliana* · Data mining · LC/MS · Metabolomic

Introduction

The tremendous development in analytical and bioinformatic methods in the last decade together with the advancement of genomics and proteomics have brought about a revolution in the manner in which biological systems are visualized and queried.

In this perspective and by taking the opportunity of the synergy created by the recent grouping of the EPGL at the University of Geneva, a new federative project for development of tools for the deconvolution and global treatment of complex bioanalytical data has started between groups involved in plant metabolite profiling, in computational chemistry, in chemometry and data mining (Fig. 1).

In a first instance efforts have been focused on the development of tools for the differential and statistical treatment of metabolomic data from crude plant extract profiling. Metabolomics has emerged relatively recently with other '-omics' technologies in biological research [1] and can be considered as the large scale analysis of metabolites that make up the metabolome. Profiling the metabolome is known to provide the most 'functional' information of the '-omics' technologies by giving a broad view of the biochemical status of an organism that can be used to monitor and assess gene function [2].

Metabolomics is however still in its infancy and the development of new strategies in this field represents important analytical and computational challenges. Indeed unlike proteomics or genomics, analysis in metabolomics is complicated by the current inability to comprehensively profile 'ALL' of the metabolome. This inability is directly related to the chemical complexity of the metabolome, the biological variance inherent in most living organisms, and the dynamic range limitations of most instrumental approaches. Thus at present no single analytical technique is capable of profiling all metabolites [3]. Furthermore a comprehensive visualization of metabolomic datasets will rely heavily upon bioinformatics. The tools needed serve to align, visualize, and differentiate components in large datasets. Individual components then need to be correlated and placed in metabolic networks or pathways. This information, together with quantitative kinetic indices, can be used to model and simulate pathways that ultimately lead to a better understanding of biological and biochemical phenomena [1].

Correspondence: Prof. J.-L. Wolfender[a]
Tel.: +41 22 379 33 85
Fax: +41 22 379 33 99
E-Mail: Jean-Luc.Wolfender@pharm.unige.ch
[a]School of Pharmaceutical Sciences
Ecole de Pharmacie Genève-Lausanne
Laboratoire de Pharmacognosie et Phytochimie
University of Geneva
Quai Ernest-Ansermet 30
CH-1211 Geneva 4
[b]LCT-Pharmacochimie
[c]LCAP-Méthodologie

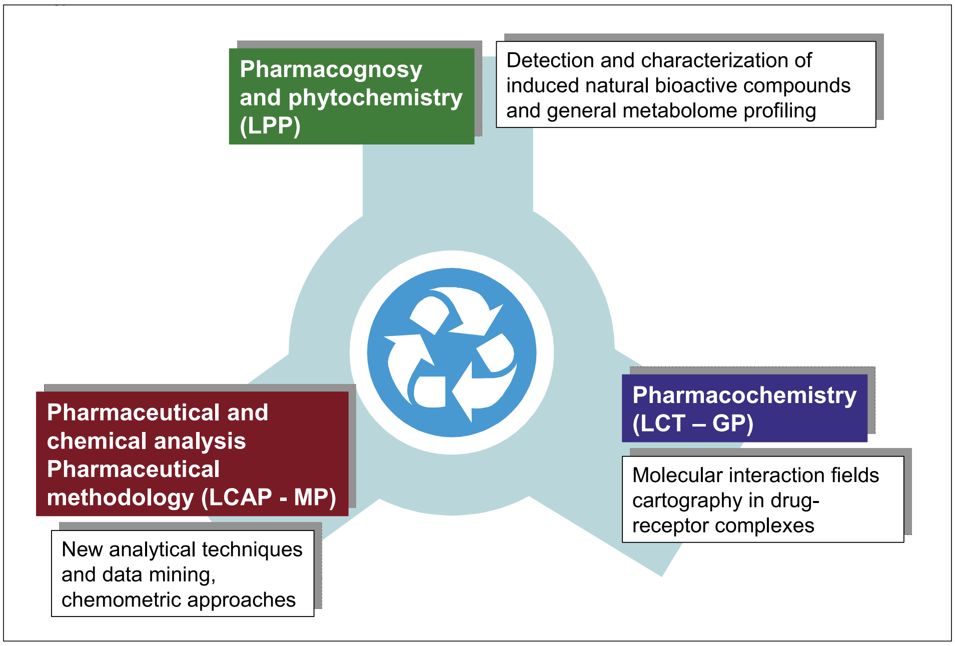
Pharmacognosy and phytochemistry (LPP)
Detection and characterization of induced natural bioactive compounds and general metabolome profiling
Pharmaceutical and chemical analysis Pharmaceutical methodology (LCAP - MP)
New analytical techniques and data mining, chemometric approaches
Pharmacochemistry (LCT – GP)
Molecular interaction fields cartography in drug-receptor complexes

Fig. 1. Synergy between three laboratories created by the recent grouping of the EPGL

Metabolomics for Seeking Biochemical Differences in Stressed Plants: The Effect of Wounding in *Arabidopsis thaliana* as a Case Study

A first project directed to the search for relevant defense induced compounds in stressed plants was chosen to establish a metabolomic strategy aimed at obtaining a global estimation of the effects of stress on metabolite expression and detailed detection of discrete induced bioactive compounds resulting from this effect. *Arabidopsis thaliana* (ecotype Columbia) was chosen as a model since total genome sequencing [4] make this plant a key target for such types of analysis and extensive data on gene expression during pathogenesis and wounding are available [5] (Fig. 2). This research is indeed motivated by the fact that based on the differential gene expression results, the induction of different wound- and pathogen-defense genes cannot be explained only by the presence of known low molecular mass regulators. Given the complexity of disease and wound signaling [6], it is thus a good possibility that signals other than those currently known are still to be discovered [7].

Analytical Approaches for the Metabolome Survey of *A. thaliana*

At present most of the work performed in defense signaling was based on the profiling of a given family of compounds, mainly oxygenated fatty acids, with targeted analytical methods such as the gas chromatography/mass spectrometry (GC/MS) profiling of oxylipins [8]. In order to obtain the most comprehensive survey possible of the leaf metabolites composition of *A. thaliana* a non-targeted analytical approach based on complementary extraction methods and liquid chromatography/mass spectrometry (LC/MS) profiling with different ionizations has been devised. As the observation of minor modifications in putative low molecular mass regulators over an important dynamic range was necessary for this project, the maintenance of good chromatographic performance was found to be mandatory to keep a good ionization efficiency of compounds that are induced in trace levels compared to the other major secondary metabolites found in the leaves. Metabolite profiling was thus performed on long HPLC columns with high separation efficiency using broad gradients in both positive (PI) and negative (NI) atmospheric pressure chemical ionization (APCI) and electrospray (ESI) LC/MS modes. The general approach is illustrated in Fig. 3.

This approach generated a huge amount of data since, for each of the MS-based detection methods, a single plant set is described by different ordered three-dimensional datasets (separation time, m/z values and intensities). Head to head comparison of the MS datasets revealed that relevant differences in the metabolite profiles were recorded between control vs wounded leaves of *A. thaliana*.

The manual or semi-automatic comparison of these matrices is however very tedious and tools for a comprehensive statistical and efficient comparison of datasets are urgently needed for an efficient filtering of compounds that are found to be up or down regulated or induced *de novo* by the stress applied to the plants. Furthermore a global estimation of the data recorded with the different methods is required in order to estimate rapidly what information is redundant and what experiments are really complementary.

Based on this model, the present research focuses on the development of tools which enable the statistical comparisons of LC/MS datasets for visualization of statistical metabolic differences. These tools provide information that can be reduced in a comprehensive manner by multivariate treatment for characterizing plant specimens in given stress conditions and that can also be used to optimize the analytical procedure towards more informative relevant experiments for a given problematic (Fig. 4).

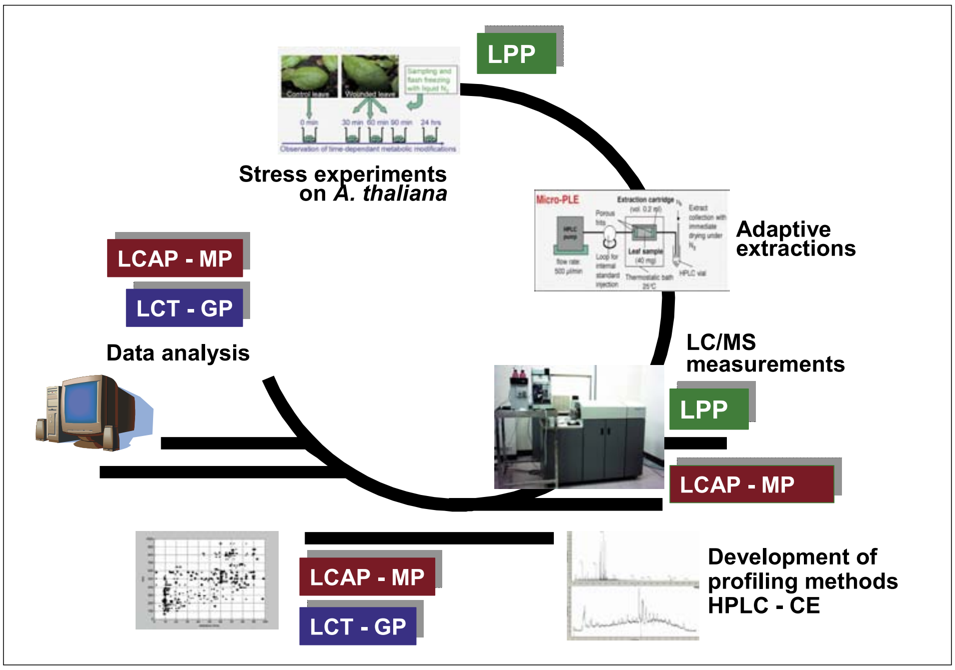

Fig. 2. Development of integrated bioanalytical strategies

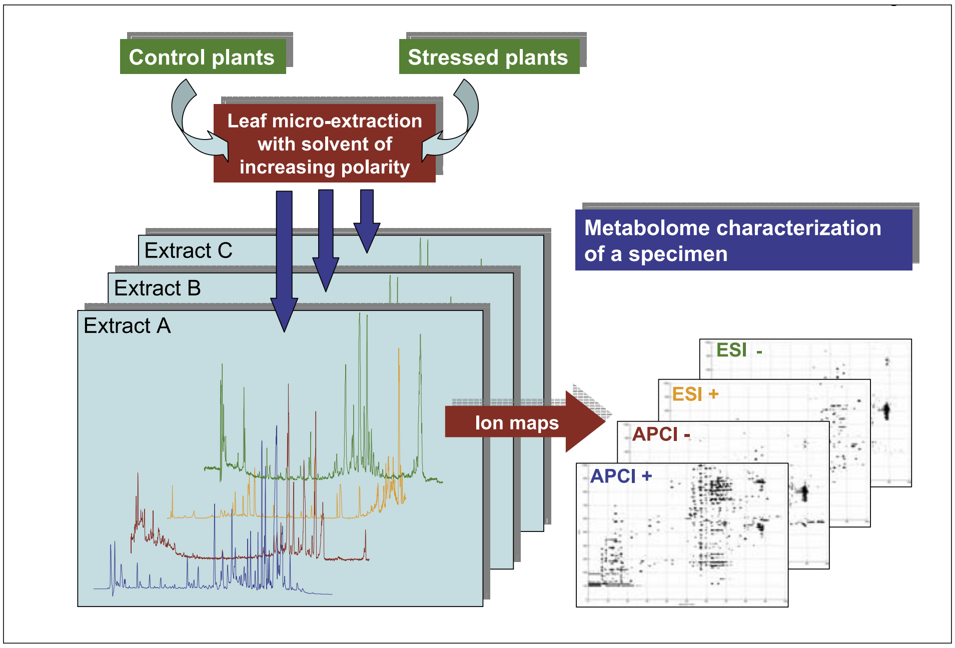
Control plants
Stressed plants
Leaf micro-extraction with solvent of increasing polarity
Metabolome characterization of a specimen
Extract C
Extract B
Extract A
Ion maps
ESI -
ESI +
APCI -
APCI +

Fig. 3. General analytical approach for metabolome profiling. Different plant specimens are extracted independently with micro extraction methods. Extracts are submitted to complementary LC/ES-MS and LC/APCI-MS analyses in both positive and negative ion modes. Metabolite profiling is provided by a set of complementary ion maps.

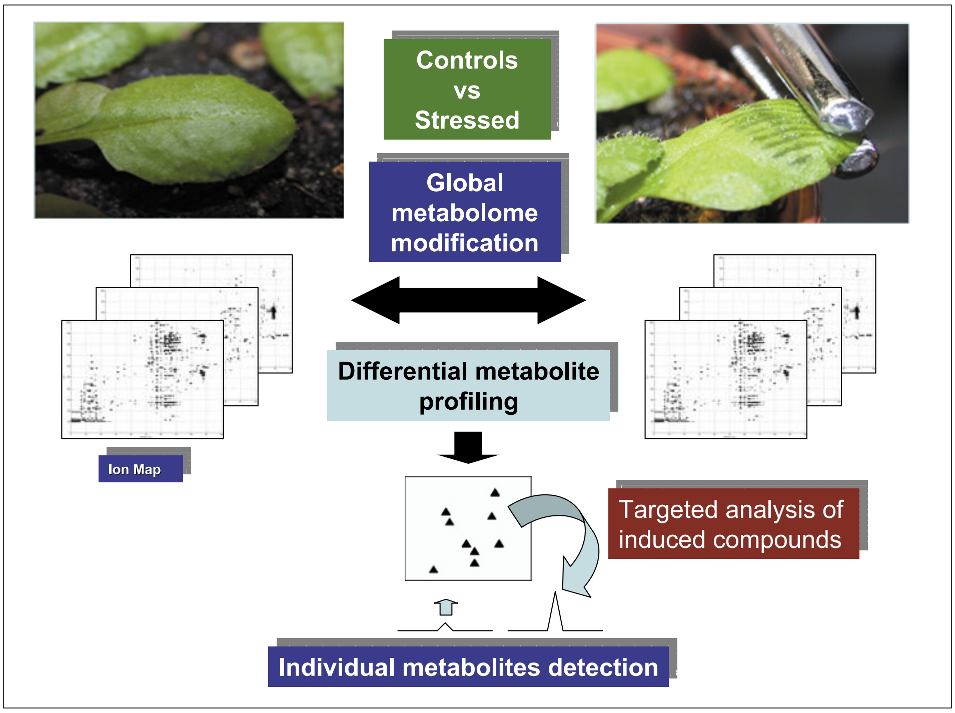
Controls vs Stressed
Global metabolome modification
Differential metabolite profiling
Ion Map
Targeted analysis of induced compounds
Individual metabolites detection

Fig. 4. Scheme for the data mining treatment of ion maps. Based on the sets of ion map recorded according to the procedure described in Fig. 3, the data can be analysed by PCA clustering for assessing global differences or for individual metabolites detection by statistic differential comparison between sets of ion maps.

Global Integration of the Results with Data Mining Tools

For the visualization of the datasets, the approach consists in parsing the MS raw data into ion maps, noise/background reduction, filtering and normalization. As each individual is characterized by complex ion maps, Principal Components Analysis (PCA) has been selected for a simplified graphical representation. Prior to PCA, a vectorization procedure is applied to each map. The vectorization is either performed vertically, corresponding to the Total Ion Chromatogram (TIC) or horizontally, equivalent to the overall Mass Composition (MC) of the extract (sum of all intensities along the time for a given m/z). Clustering techniques such as hierarchical clustering analysis (HCA) were further applied for data analysis.

As displayed in Fig. 5, a discrimination between wounded and control leaf of *A. thaliana* (NI LC/APCI MS of the lipophilic extract) over two different days was obtained by PCA and HCA. While intra-day clustering between wounded and control samples could be easily achieved, the main difficulty of this holistic LC/MS approach is related to day-to-day method variability. Thus, differences between plants might be hidden by instrumental signal intensity variations, leading to erroneous data clustering. To overcome this problem, ion maps normalization procedures including reference signals used as landmarks are currently under investigation.

Treatment of these global differences will give a good view of the parameters that discriminate plant sets. It can also be very helpful in understanding the global structure of the data such as specific metabolite tracking and observation of pattern of correlated modifications related with external parameters (*e.g.* extraction solvent, ionization techniques, type of chromatography, *etc.*). An exhaustive profiling of 'ALL' metabolites with different detection methods is very demanding in terms of analysis time and instrumental resources. Therefore, relevant complementary experimental conditions could be also selected based on this approach for a rational and comprehensive visualization of the metabolome.

Selection of determinant ion maps based on this treatment has thus important chances to reveal significant differences that would not be easily detected just based on simple comparisons. With optimized key experiments, new target compounds produced by stress induction would be efficiently localized as discriminant scores in PCA.

For a detailed comparison of differences between samples and precise localization of metabolites involved in a given plant response with a selected analytical procedure, generation of ion map of statistical difference (Student or Anova test) could be achieved. An example of this data treatment performed on the NI LC/APCI-MS analysis of control versus stressed leaf sample sets (wounding with forceps) is shown in Fig. 6. On this map, significant modifications in metabolites concentration after injury caused by up- or down-regulation as well as *de novo* expression were evidenced. For example, a closer look into the oxygenated fatty acids region (t_r: 20–50 min) revealed a peak (m/z 209) not detected in controls and present in all wounding samples. Identification of this compound revealed that it corresponded to jasmonic acid (JA), a referenced wound inducible signal, usually only evidenced with targeted methods [8].

These preliminary results demonstrate that peaks statistically induced *de novo* in wounded leaves can be efficiently tracked by this visualization method.

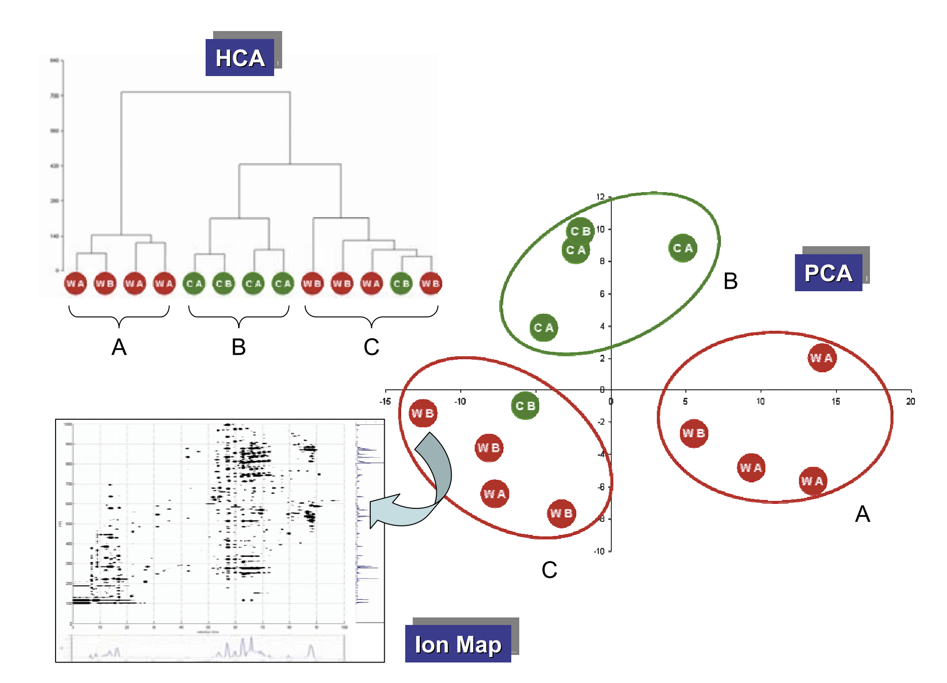

Fig. 5 Multivariate treatment of the ion map information. Clustering is provided by vectorization of the individual ion maps. Group of control samples obtained within two different days of analysis (CA) and (CB) are efficiently differentiated from wounded samples (WA) and (WB).

Conclusion

With the synergy created by the recent grouping of the EPGL at the University of Geneva, the investigation of complex biological system, such as the study of global metabolite variations in living organisms, can be tackled with an efficient multidisciplinary approach at the interface of phytochemistry, analytical biochemistry, chemometric and computational sciences.

The improvement of the LC/MS profiling methods in terms of resolution, reproducibility, choice of optimal ionization conditions give an overview of the global changes that occur at the metabolome level. Because of the important chemical diversity of natural products and the convolute nature of biological matrices, complementary analytical procedures need to be employed. The determination of all constituents, together with adapted chemometric tools, allow the efficient localization of relevant induced metabolites for a given biochemical response. Thus, the chances to interpret a plant metabolome would gradually increase with the thorough evaluation of the complete metabolome profile. The multivariate treatment of this complex array of data gives also the possibility to evaluate redundant information, to select orthogonal analytical approaches and therefore, to better discriminate individuals. Thus, for a given problematic, a selection of the most relevant procedures and stress induction experiments would be facilitated, considerably accelerating the investigations.

The developed approach has already been demonstrated to be efficient for the localization of unknown induced low molecular mass compounds in wounded leaves of *A. thaliana* by the help of differential display (statistical difference ion maps). Based on these results, an LC/MS triggered collection of the induced compounds has been performed and the complete structural determination at the microgram level is underway with highly sensitive microflow capillary LC/NMR methods [9]. The compounds of interest will be tested for their potential for defense gene expression (collaboration with Prof. E.E. Farmer, Plant Molecular Biology, UNIL, BB, CH-1015 Lausanne). Therefore, the possibility to efficiently analyze metabolome changes in given stress and time conditions would shed more light on data observed at the genomic level upon multivariate analysis of differential gene expression results.

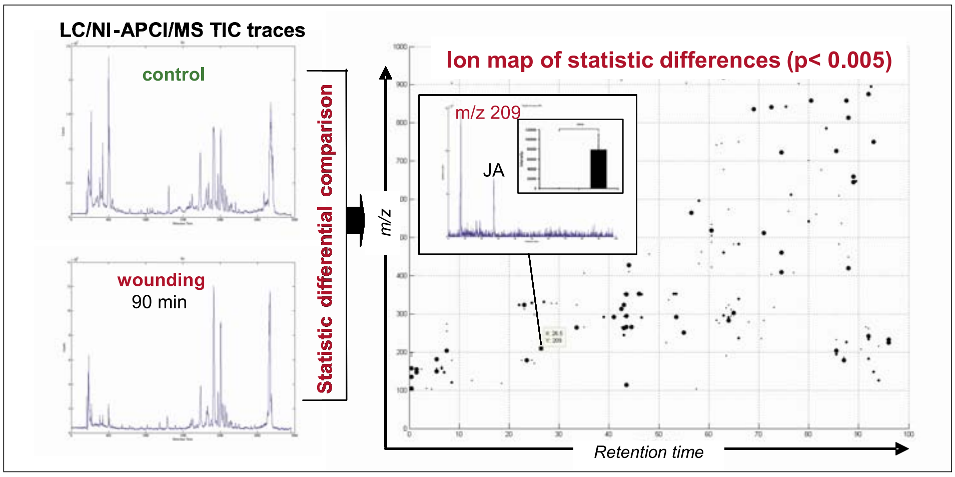

Fig 6. The comparison of sets of LC/MS data from control versus wounded plants provide ion map of statistic differences. On this type of maps big dots correspond to significant differences (p>0.005) between plant sets. As shown in the inset the dot recorded at m/z 209 with a retention time of 26 min correspond to jasmonic acid JA, which was thus found to be significantly induced 90 minutes after wounding.

Acknowledgements

We thank Prof. Edward Farmer for guidance in experimental design related to stress induced experiments. Financial support was provided by the Swiss National Science Foundation (Grant no. 205320-100396 /1 to J.-L. Wolfender)

Received: April 19, 2005

[1] L.W. Sumner, P. Mendes, R.A. Dixon, *Phytochemistry* **2003**, *62*, 817.

[2] O. Fiehn, J. Kopka, P. Dörmann, T. Altmann, R.N. Trethewey, *Nature Biotechnol.* **2000**, 18.

[3] O. Fiehn, *Plant Molecular Biology* **2002,** *48,* 155.

[4] T. Arabidopsis Genome Initiative, *Nature* **2000,** *408/14,* 796.

[5] P. Reymond, H. Weber, M. Damond, E.E. Farmer, *Plant Cell* **2000,** *12,* 707.

[6] B.P. Thomma, I.A. Penninckx, W.F. Broekaert, B.P. Cammue, *Current Opinion in Immunology* **2001,** *13,* 63.

[7] E.E. Farmer, E. Almeras, V. Krishnamurthy, *Current Opinion in Plant Biology* **2003,** *6,* 1.

[8] H. Weber, B.A. Vick, E.E. Farmer, *Proceedings of the National Academy of Sciences of the United States of America* **1997,** *94,* 10473.

[9] D.L. Olson, J.A. Norcross, M. O'Neil-Johnson, P.F. Molitor, D.J. Detlefsen, A.G. Wilson, T.L. Peck, *Analytical Chemistry* **2004,** *76,* 2966.