

An integrated approach to functional genomics and bioinformatics in a model legume

PI: Pedro Mendes, Virginia Bioinformatics Institute (VBI), Virginia Tech (0477), 1750 Kraft Drive, Suite 1100, Blacksburg, VA 24061

Co-PI: Richard A. Dixon, Plant Biology Division, Samuel Roberts Noble Foundation (SRNF), 2510 Sam Noble Parkway, Ardmore, OK 73401

Collaborator: Lloyd Sumner (SRNF)

Collaborator: Gregory D. May (SRNF)

Collaborator: Jennifer Weller (VBI)

Collaborator: Tim Smith, Department of Physical Sciences, South Eastern Oklahoma State University (SOSU), P.O. Box 4025, Durant, OK 74701

C. PROJECT DESCRIPTION

We propose an integrated functional genomics approach to study the relationships between gene expression, protein levels and metabolites in the model legume *Medicago truncatula*, with particular emphasis on natural product pathways. We will accomplish this by studying elicited cell cultures through an arsenal of analytical techniques, and developing software to integrate the diverse data sets and explain them through quantitative predictive models. This study will encompass measuring quantitative information for mRNA, protein, and metabolite pools and will include development and implementation of a comprehensive bioinformatics system to archive and analyze the results of these studies. *Medicago truncatula* provides an excellent model system for a true molecular genetic dissection of the pathways leading to agronomically important bioactive natural products such as flavonoids, isoflavonoids and triterpenes, and these pathways will be a major focus of the proposal. At the same time, the bioinformatics resources that will be developed will be of general utility for integrating metabolic profiling data with gene expression and proteomic data and for reverse engineering genetic and metabolic networks from functional genomics data.

The PIs on this project share a common interest in the cellular regulation of plant biochemistry and believe that this can only truly be done by combining detailed observations from gene expression and metabolism. The team that has been assembled here combines all the expertise necessary to do so, ranging from diverse fields such as plant cell culture, cDNA microarrays, proteomics, metabolic profiling, databases, numerical analysis and metabolic modeling. In addition, most of the technology required to generate these data sets already exists in the participating laboratories.

The current request is for support for personnel to perform the experiments and to research and implement the algorithms to relate the data to each other. A request is also made for matching funds to allow the purchase of a Q-TOF mass spectrometer equipped with capillary electrophoresis for peptide sequencing.

Functional Genomics - Functional biology of a cell

Functional genomics seeks to determine the cellular and organismal functions of genes and the manner in which their products interact to maintain those functions. Determining the first level of gene expression consists of monitoring RNA species and their concentrations, but, to truly understand the control of cellular function, it is necessary to extend the analysis further to protein and metabolite concentrations. It is only by considering these three levels in a biological system in which gene expression can be flexibly manipulated that we can fully address the relationships between gene expression and its downstream consequences for the cell.

The biological system that will be studied in the present proposal is a cell culture. This choice stems from technical reasons, and also the need to minimize the number of variables involved. In the

living plant the interaction of RNA, protein and metabolites would affect phenotype, including morphology and physiology. While in the model system here proposed such effects are lost, the authors see great value in utilizing a model system for establishing methods that can later be applied to more complex systems.

A major goal of this project is the generation of integrated data sets leading to development of bioinformatic tools for true functional genomic analysis. The metabolic systems to be analyzed, and the approaches to be taken, derive logically from the current research of the PI (systems approaches to functional genomics including “reverse engineering of genetic and biochemical networks”), and Co-PI (biochemistry, functional genomics and metabolic engineering of natural products in legumes), and the results will synergistically enhance both programs. At the same time, analysis of this model system will allow us to address a number of important and specific biological questions. These include:

1. *Which genes are re-programmed at the transcriptional level in response to biotic and abiotic elicitation?*
2. *What are the kinetics of transcriptional, translational and metabolic changes?* It has become a tradition in many scientific publications to indicate that metabolism is faster than gene expression. Despite this there are theoretical arguments that indicate such time scale separation can be destroyed by feedback from the metabolic level to the translation apparatus. The proposed research will directly measure these time scales and will contribute to clarifying this issue. This is an important open issue for all functional genomics studies since if all three levels (mRNA, proteins and metabolites) operate on comparable time scales then microarray measurements, in isolation, are much harder to interpret than has been assumed until now.
3. *Are there specific patterns of expression for sets of genes, and what is the physiological significance of these patterns based on the predicted functions of these genes?*
4. *Are changes in gene expression necessarily followed by corresponding changes at the protein and metabolite levels?* For example, elicitation of alfalfa cell cultures leads to strong transcriptional activation of genes encoding lignin *O*-methyltransferases, but the transcripts do not accumulate in the polysomal RNA fraction and there is therefore no corresponding increase in enzyme activities or accumulation of lignin (Ni *et al.*, 1996b). Comparison of transcript levels to protein levels will enable us to determine which genes are activated but not translated in elicited cells.
5. *Are changes in gene expression causally linked to changes in metabolite levels?* Phenylpropanoid/isoflavonoid pathway intermediates can act as regulators of their biosynthetic pathways at the genetic and enzymatic levels (Dixon *et al.*, 1980; Gerrish *et al.*, 1985; Loake *et al.*, 1991). Such relationships can now be fully explored by comparison of global gene expression patterns with metabolic profiles.

A. The biological system

A.1. Background and rationale

Our choice of a biological system was dictated by several criteria. The plant species must be relevant to important agronomic crops, and amenable to genomic and genetic analysis. The system must be simple, reproducible, homogeneous with respect to cell type, inducible to perturb gene expression, and amenable to extraction of RNA, proteins and metabolites with minimal difficulties. In-depth knowledge of at least one inducible metabolic pathway in the organism of choice is also helpful, as an "internal control pathway" for optimization of induction conditions and to serve as a subject for a more detailed analysis of the relation between gene/protein expression and metabolism.

A.1.a. The organism

Medicago truncatula is closely related to the world's major forage legume, alfalfa, and has been chosen as a model species for genomic studies in view of its small genome, fast generation time and high transformation efficiency (Cook 1999; Trieu *et al.*, 2000). Genes from *M. truncatula* share very high sequence identity to their counterparts from alfalfa (e.g. 98.7 and 99.1% at the amino acid levels for

isoflavone reductase, and vestitone reductase, respectively), so it serves as an excellent, genetically tractable, model for alfalfa. Studies on synteny relationships in the laboratory of Dr Doug Cook (U.C. Davis) are establishing links between *M. truncatula*, alfalfa, and pea, as well as Arabidopsis, for which the first complete plant genome sequence will soon be available.

As a legume, and unlike the most studied genetic model plant, Arabidopsis, *M. truncatula* establishes symbiotic relationships with nitrogen fixing Rhizobia. Roots of *M. truncatula* are also colonized by beneficial arbuscular mycorrhizal fungi (Harrison & Dixon, 1993). Importantly for the present proposal, the complex interactions of legumes with microorganisms have resulted in the evolution of a rich variety of natural product biosynthetic pathways impacting both mutualistic and disease/defense interactions. Of these, the isoflavonoid pathway, which is not present in Arabidopsis, leads to nodulation gene inducers and repressors, pterocarpan phytoalexins involved in host disease resistance, and isoflavones with anticancer and other health promoting effects for humans. This pathway has been well characterized in alfalfa, and in other legumes such as soybean and chickpea, at the metabolic, enzymatic and genetic levels (Paiva *et al.*, 1994; Dixon *et al.*, 1995; Dixon, 1999), and serves as the "internal control pathway" for the proposed studies.

M. truncatula is currently the subject of two major genomics initiatives in the United States. A NSF funded program is producing ESTs, and performing map-based cloning of symbiotic genes, comparative genomics and BAC survey sequencing. The internally funded program at SRNF has established a large-scale EST database, and is utilizing T-DNA activation tagging as well as proteomic analysis and metabolite profiling to provide complete coverage of functional genomics in this species.

A.1.b. *The plant cell culture system*

Elicitor-induced plant cell cultures have been used for many years to study defense gene activation and associated metabolic changes (Dixon, 1980; Schmelzer *et al.*, 1984; Barz & Mackenbrock, 1994). Despite being simplified model systems, studies with such cell cultures have generated much of the available information on biochemical responses of plants to biotic stress. Many of the responses have been shown to occur, with similar kinetics, during natural infections (Schmelzer *et al.*, 1984; Latune-Dada *et al.*, 1987; Kombrink *et al.*, 1990; Paiva *et al.*, 1994; Batz *et al.*, 1998). The system therefore has relevance to more complex biological states. The requirement that we establish baselines for noise in these types of experiments means that we need relatively large samples of a single organism and cell type for which reproducible perturbation and sampling are achievable. This has been demonstrated for plant cell cultures (Dixon, 1980; Chappell & Hahlbrock, 1984).

A.1.c. *The isoflavonoid pathway*

M. truncatula cell suspension cultures respond to microbial elicitation in a similar manner to alfalfa cell cultures by accumulating the isoflavonoid phytoalexin medicarpin (Fig.1). We have previously characterized genes from alfalfa, *M. truncatula*, and/or soybean that encode eleven of the enzymes involved in the formation of medicarpin from L-phenylalanine (Fig.1, genes listed in the figure legend). Because these genes are already characterized microarray technology and MALDI-TOF mass spectrometry may be used to assay, respectively, mRNA and protein levels for nearly all the enzymes of this pathway from the inception of the project. In addition, we have developed routine methods for HPLC separation of *Medicago* isoflavonoids (Kessmann *et al.*, 1990; Harrison & Dixon, 1993) (see Section B.2.d). Finally, there is a considerable amount of existing knowledge on the substrate specificity and kinetic properties of many of the isoflavonoid pathway enzymes (Dixon, 1999), which will facilitate modeling of pathway fluxes.

Using the isoflavonoid pathway as our first target will allow us to control our experimental procedures so that we can subsequently acquire reliable data about less-well characterized, or previously uncharacterized, pathways.

A.1.d. *Relevance of elicitors*

Three different elicitors, two biotic and one abiotic, will be used in the proposed studies. These are: purified yeast elicitor, for which a great deal of data already exists; methyl jasmonate, a putative signal transduction component in elicitation responses; and UV irradiation, a known environmental effector.

Purified yeast elicitor consists of cell wall polysaccharides composed entirely of mannose (Schumacher *et al.*, 1987). This elicitor has been shown to strongly induce the isoflavonoid pathway and associated primary metabolism, such as the pentose phosphate pathway, in alfalfa (Paiva *et al.*, 1994; Fahrendorf *et al.*, 1995). It is also a good inducer of isoflavonoids in *M. truncatula* (see Section B.2.d). The same preparation has been used for elicitation of other pathways in plant cell cultures, including benzophenanthridine alkaloids (Schumacher *et al.*, 1987; Funk *et al.*, 1987), indole alkaloids (Gundlach *et al.*, 1992), prenylated isoflavonoids (Funk *et al.*, 1987), hydroxycinnamoyl amides (Muhlenbeck *et al.*, 1996), and ethylene (Grosskopf *et al.*, 1990).

Methyl jasmonate (MeJa) was shown to induce natural product accumulation in all 36 species tested in a cell suspension culture survey (Gundlach *et al.*, 1992). Levels of MeJa increase in a range of plant cell cultures in response to yeast elicitor (Gundlach *et al.*, 1992). It was therefore proposed that MeJa is an obligate signal transduction component for elicitor-induced phytoalexin responses, irrespective of the nature of the pathway being induced (Gundlach *et al.*, 1992; Mueller *et al.*, 1993). However, MeJa and yeast mannan elicitor do not always induce the same compounds; for example, in soybean cell suspension cultures, yeast elicitor induces the pterocarpan glyceollin whereas MeJa induces the isoflavone genistein (Gundlach *et al.*, 1992). Biochemical and genetic studies with intact plants have indicated that MeJa may be involved in activation of a specific sub-set of genes related to herbivory-induced wound responses that may be affected by negative cross-talk with anti-microbial defense responses (Doares *et al.*, 1995; Felton *et al.*, 1999). A recent study demonstrated subtle differences in transcriptional and post-transcriptional control of terpenoid phytoalexin biosynthesis in tobacco cell cultures in response to fungal elicitor and MeJa (Mandujano Chavez *et al.*, 2000). A global comparison of gene expression patterns in response to yeast mannan elicitor and MeJa will provide the necessary information to assess the extent of independence between the pathways they induce and help identify genes involved in them.

Previous studies in parsley cell cultures and protoplasts have shown that UV exposure induces the same core phenylpropanoid pathway genes as fungal elicitor, although with different kinetics, and different downstream branch pathway genes lead to accumulation of UV-protective flavonoid pigments (Chappell & Hahlbrock, 1984; Dangl *et al.*, 1987; Hahlbrock & Scheel, 1989). The UV-induced phenylpropanoid pathway shows different kinetics of induction of core pathway genes such as phenylalanine ammonia-lyase (PAL), compared to fungal elicited cells, which may reflect induction of specific gene family members. In alfalfa, there are at least six PAL genes and more than eight chalcone synthase (CHS) genes. It is not known whether these genes have specific functions, for example in the formation of UV protective flavonoids or defense-related isoflavonoids. Comparison of the induction kinetics of the members of multigene families encoding PAL, coumarate CoA ligase (4CL) and chalcone synthase (CHS) with quantitative changes in metabolites in response to yeast elicitor, MeJa and UV, will help to address this important question.

UV radiation in sunlight creates two main DNA damage products in the form of cyclobutane pyrimidine dimers (CPD) and pyrimidine-(6-4')-pyrimidone photoproducts. These types of DNA damage have been shown to directly inhibit DNA replication and transcription in prokaryotic and mammalian systems (Protic-Sabljić *et al.*, 1986; Mitchell *et al.*, 1989; Donahue *et al.*, 1994). A significant emphasis of studies of plant DNA repair mechanisms has focused on UV-induced damage, which is due in large part to concerns associated with the deterioration of the stratospheric ozone layer (Vonarx *et al.*, 1998). Repair of UV-induced damage occurs through mechanisms that at least include light-dependent photoreactivation (mediated by photolyases), translesion synthesis, and nucleotide excision repair (NER). In addition to strategies of DNA-damage/repair/tolerance, plants have evolved shielding mechanisms (Pang *et al.*, 1993). For example, flavonoids such as flavones, isoflavonoids and anthocyanins accumulate in the vacuoles of the epidermal layers of alfalfa in response to UV-B exposure (Takayanagi *et al.*, 1994). Mutants unable to synthesize these compounds have decreased resistance to UV-induced damage,

illustrating the importance of these compounds in protection against UV-exposure (Li *et al.*, 1993; Lois, 1994).

Conconi *et al.* (1996) reported similarities between wounding and UV exposure in activating systemic wound-response genes in tomato leaves. They suggested that UV irradiation may lead to perturbation of membranes and/or the activation of lipase activity leading to the release of linolenic acid thereby initiating the MeJa signal transduction pathway to activate wound inducible genes. Thus there may be points of crossover among all three elicitors as well as points of distinction. Studies of global gene and protein expression changes in UV-exposed cells are likely to lead to discovery of new genes involved in DNA repair (a major research area in Dr Greg May's laboratory at SRNF) and chemical UV defense mechanisms.

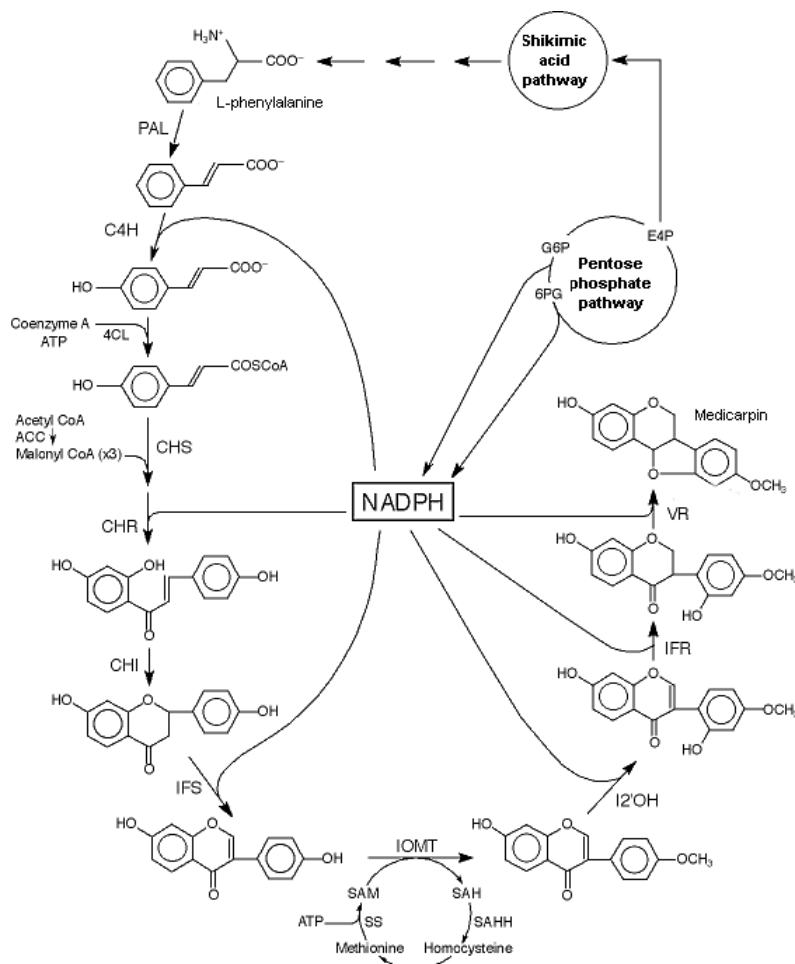


Fig. 1 - Biosynthesis of isoflavonoids in *Medicago*. Full sequences for the following gene products are known: L-phenylalanine ammonia-lyase (PAL, Gowri *et al.*, 1991); cinnamate 4-hydroxylase (C4H, Fahrendorf & Dixon, 1993); acetyl CoA carboxylase (ACC, Shorrosh *et al.*, 1994); chalcone synthase (CHS, Junghans *et al.*, 1993); chalcone reductase (CHR, Balance & Dixon, 1994); chalcone isomerase (CHI, Maxwell *et al.*, 1993); isoflavone synthase (IFS, Steele *et al.*, 1999); isoflavone O-methyltransferase (IOMT, He *et al.*, 1998); isoflavone 2'-hydroxylase (I2'OH, C.L. Steele & R.A. Dixon, unpublished results); isoflavone reductase (IFR, Paiva *et al.*, 1991); and vestitone reductase (VR, Guo & Paiva, 1995). EST sequences representing

isoforms of several of these genes are also available in the SRNF *Medicago* EST database.

A.1.e. Specificity of elicitation

Historically, studies of responses to elicitation have focused on single biochemical pathways, in large part because of the tools available to scientists. Therefore little is known concerning the extent, specificity, and level of genetic re-programming following elicitation. One hypothesis is that a limited range of "elicitor response" cis-elements is present in the regulatory regions of a discreet subset of genes that are elicitor responsive. Several different cis-elements have now been identified by functional analysis of defense response gene promoters (e.g. van de Löcht *et al.*, 1990; Yu *et al.*, 1993; Raventos *et al.*, 1995; Yang *et al.*, 1999). However, this does not address the question of how many genes change their expression in response to elicitation. An indication that this number might be large can be deduced from the results of experiments using parsley cells (Batz *et al.*, 1998). Computer simulations of models integrating gene expression and metabolism lead to the conclusion that perturbations that specifically target one gene end up affecting other apparently unrelated genes (i.e. without common upstream regulatory sequences), the perturbation being transmitted to the latter by the changes at the metabolic level (Mendes, 1999).

Exposure of parsley cells to a defined peptide elicitor from the non-host fungus *Phytophthora sojae* leads to many and varied changes in gene expression. Transcript levels increase for enzymes involved in the furanocoumarin phytoalexin pathway, the formation of cell wall amines and phenolics, the SAM cycle and ethylene formation, the shikimate pathway, glycolysis, the pentose phosphate pathway, and fatty acid biosynthesis (Kawallek *et al.*, 1992; Batz *et al.*, 1998). There are also decreases in transcript levels for a number of genes involved in photosynthesis and the cell cycle. In most cases, however, it is not known from direct measurements whether transcriptional activation of these genes has any metabolic consequence; this has only been inferred from assumptions of metabolic associations between pathways, not determined directly by measurement of protein and metabolite pools. We propose to fill this gap with the present study.

A.1.f. Discovering novel genes for defense and natural product biosynthesis

Elicitor-induced mRNAs that are translationally active in the cell cultures are candidates for defense response genes. In addition to the enzymes of the phytoalexin response, many other defense-related proteins are produced in infected or elicited cells. These include various classes of pathogenesis-related proteins (Somssich *et al.*, 1986), antimicrobial proteins such as defensins (Broekaert *et al.*, 1995), components of the oxidative burst machinery (Auh & Murphy, 1995; Desikan *et al.*, 1996), among others. Detailed gene expression profiling should reveal new members of these gene families as well as completely novel genes, which might be of value for crop improvement.

Zenk (1991) described plant cell cultures as a "pot of gold" for discovery of the enzymes of plant secondary metabolism. Alfalfa and *M. truncatula* are rich sources of an agronomically important class of natural products, the triterpene saponins. These compounds, of which medicagenic acid is a major aglycone in *Medicago* species, have anti-insect activity (Tava & Odoardi, 1996), antinutritive and anti-feedant effects for some monogastric animals (Small, 1996), are hemolytic (Oleszek, 1996) and, from some plant species, have potent anticancer activity (C.A. Arntzen & J. Gutterman, personal communication). Little is known of the later stages of their biosynthesis from the C30 precursor squalene. We have been able to extensively profile *M. truncatula* saponins, including glyco-conjugates of medicagenic acid, hederagenin, bayogenin, and soyasaponin I, by HPLC/MS (see Section B.2.d).

Mining our *M. truncatula* EST database has revealed candidate clones for squalene synthase, squalene epoxidase and β -amyrin synthase, the first three enzymes of triterpene saponin biosynthesis, and these genes are currently being functionally characterized. Comparative analysis of gene expression and metabolite profiles, in parallel to screening for saponins in T-DNA activation tagged lines of *M. truncatula* (ongoing in a parallel project at SRNF) will provide a valuable new approach for gene

discovery in the saponin pathway. A similar strategy for gene discovery in the coumestan pathway, a branch of isoflavonoid biosynthesis yet to be explored at the molecular level, will be pursued in a parallel project in the laboratory of Dr. Nancy Paiva at SRNF.

B. Experimental design.

B.1. Biology - establishment and elicitation of cell suspension cultures of *Medicago truncatula*

Cell suspension cultures of *M. truncatula* cv Jemalong A-17 will initially be grown in shaker flasks in modified Schenk and Hildebrandt medium containing 2,4-D, *p*-chlorophenoxyacetic acid and kinetin (Franklin & Dixon, 1994). These conditions lead to good growth with production of medicarpin following elicitation. Elicitation conditions will be optimized in a set of preliminary experiments by varying auxin/cytokinin levels and elicitor concentration (Dixon *et al.*, 1981). Yeast elicitor will be tested at concentrations between 20 µg and 2 mg glucose equivalents/mL culture (Schumacher *et al.*, 1987); the optimum for medicarpin induction in alfalfa cell cultures is around 50 µg/mL. MeJa will be tested between 1 and 100 µM (Gundlach *et al.*, 1992). We will then determine the optimum stage in the growth phase of the cultures for elicitation (Kombrink & Hahlbrock, 1985). This can be monitored subsequently by measuring changes in the conductivity of the culture medium (Carrier *et al.*, 1990). The conditions for optimum elicitation are known for alfalfa cell suspension cultures, and are expected to be similar for *M. truncatula*. Once conditions are optimized for yeast elicitor-mediated isoflavonoid induction, we will check, using low-density gene expression microarrays and profiling of flavonoid/isoflavonoid metabolites, whether the same conditions are suitable for induction with MeJa and UV. Ideally we will use identical elicitation conditions for yeast elicitor, MeJa and UV, providing it is possible to observe gene expression changes with all three treatments with the same culture conditions and growth stage.

Accumulation of natural products usually occurs within the first 48 h of elicitation, with most of the major changes in gene expression occurring during the first 12 h, and many within the first 30 min (Ni *et al.*, 1996a; Batz *et al.*, 1998). We will sample elicited cell cultures, in triplicate, at 0, 5, 15, 30, and 45 min, 1, 2, 3, 4, 6, 8, 10, 12, 18, 24, 30, 36, 42 and 48 hours post-elicitation. Control cultures will be harvested at 5, and 30 min, 1, 3, 6, 12, 18, 36 and 48 hours after treatment with water. We will need approximately 10 g fresh weight of cells for each sample for gene expression, proteomic and metabolite analyses. This amounts to approximately 1 kg of cells for a complete induction experiment. We will weigh (wet fresh weight) to equally divide pooled inoculum into the needed number of flasks so that each flask starts the same, allow the cells to grow for 3-5 days, then treat and harvest at the appropriate time. This minimizes the stress due to handling immediately before elicitation, and speeds up the harvest process, which involves pouring the contents of the flask onto a nylon mesh filter, putting under suction for 5-10 seconds, and then freezing in liquid nitrogen.

Experimental conditions for UV-elicitation of *M. truncatula* cell suspension cultures have previously been established at SRNF in the May laboratory. Briefly, actively dividing cell suspensions are plated onto solid medium and are exposed to 25,000 µJ cm⁻² of UV at a peak wavelength of 254 nm. Samples are harvested at the same times as for yeast elicitor and MeJa treatments. An alternative treatment includes a second UV exposure 30 minutes after the initial treatment. Preliminary evidence suggests that both treatments result in differential patterns of protein accumulation on SDS-PAGE gels in comparison to the untreated controls. At higher doses of UV exposure, transcript levels of a DNA repair protein mRNA begin to diminish. Conconi *et al.* (1996) similarly reported diminished levels of protease inhibitor gene induction in tomato leaves overexposed to UV.

B.2. Technologies and data sets

B.2.a. *Medicago truncatula* ESTs

An internally funded core program on *Medicago* genomics has been established at SRNF. The program is taking a global approach to studying the physiology and biochemistry of *M. truncatula*, with special emphasis on natural product biosynthetic pathways, mycorrhizal interactions and factors impacting forage quality. Approaches include large-scale EST sequencing, gene expression profiling, the

generation of *M. truncatula* activation-tagged and enhancer trap insertion mutants, high-throughput metabolic profiling, and proteome studies. These studies complement, rather than overlap with, the current NSF-funded *Medicago truncatula* genome project. To house these research activities within the Plant Biology Division, SRNF's Board of Trustees recently approved funding for the construction of a new, three-story research complex to be built on the SRNF campus. Construction is scheduled to be completed in the fall of 2001.

The Medicago Genome Initiative (MGI) is a publicly available database of SRNF-generated *M. truncatula* EST sequences (Bell *et al.*, 2000) currently stored at NCGR (National Center for Genome Resources, Santa Fe, New Mexico). MGI's software system consists of three interacting sub-systems, a relational database for storage of the sequence data and the results of its analysis, an automated analysis pipeline that performs the analyses, and a user interface, which presents a variety of views of the data to the researcher. The user interface can be modified and developed somewhat independently of the other sub-systems. This allows considerable flexibility in implementing novel ways of analyzing and presenting data in response to the needs of the user community. The *Medicago truncatula* Gene Index at TIGR (<http://www.tigr.org/tdb/mtgi/>) is also publicly available and is inclusive of the data generated by SRNF, NSF and international *Medicago* EST programs.

Since January 2000, more than 55,000 *M. truncatula* ESTs have been characterized at SRNF. Sequence information is being generated on an ABI 3700 DNA Analyzer. A series of robotics stations fully automate sample processing prior to analyses on the ABI 3700. Separate, developmentally pooled unidirectional cDNA libraries were generated in the UniZap (StrataGene) cloning vector system. These libraries represent transcripts from roots, stems, leaves, nodulated roots, drought-stressed plantlets, elicitor-induced cell cultures (critical for the present project), insect herbivory damaged leaves and phosphate-starved leaves. Two root cDNA libraries are of particular relevance to the present project. One represents transcripts from roots harvested 4 weeks after nodulation by *Rhizobium meliloti*, and contains nodules and roots at different stages of development. The second represents non-nodulated roots harvested at 2, 8, and 16 days and 6 weeks after germination and grown in perlite wetted with a modified Hoagland's solution. Roots are a rich source of isoflavonoid and terpenoid natural products. A library from *Phoma medicaginis*-infected leaves is being constructed from tissue harvested at 0, 15 min, 30 min, and 1, 2, 3, 6, 14, 24, 48, 72, and 96 hours after inoculation. For the latter two libraries, total RNA was isolated from each time point, equal amounts were pooled, and used to prepare mRNA. A portion of the tissues used in the construction of these libraries has been subjected to HPLC analysis. The presence of formononetin (an isoflavone) and medicarpin conjugates was confirmed in the root samples, increasing with time of harvest, whereas the accumulation of medicarpin and coumestrol was observed in the infected tissues.

TraceTuner (Paracel) quality scores are used for sequence trimming. Sequences with an average quality score of less than 15 (Q15) over 20 sliding bases are removed. (Paracel recommends a Q12 cutoff for EST sequences.) Minimum insert size of 60 bp (Q15) is required for an EST to be included in the database. Insert sequences are automatically trimmed after 800 bases. We are currently experiencing a greater than 90 percent success rate using these parameters. Our latest average reads are greater than 700 bases in length. We continue to experience a relatively low level of redundancy in our libraries based on clustering results using TIGR's Assembler software (see Table 1 below). This may be due in part to the large number of libraries in the program and the shallow depth at which they have been analyzed to date.

Table 1 – MGI clustering statistics for the first 33,340 ESTs.

Source Library	Input Sequence	Non-redundant Sequences
Nodulated root	3,070	2,514
Leaf	5,730	5,177
Root	2,948	2,851

Stem	10,750	7,975
Insect herbivory	3,010	2,757
Drought	4,270	4,001
Elicited cell culture	1,510	1,459
Phosphate-starved leaves	2,052	1,946
All Combined	33,340	26,197

We will continue to sequence clones from the current libraries until approximately 10,000 ESTs are characterized from each, or redundancy issues arise. We will generate and sequence additional libraries from flower (emergence through developed seed pods), developing seed and germinating seed (imbibed seed through 5mm radical length) tissues, *Phytophthora*-challenged leaves and alfalfa mosaic virus-infected tissues.

Unique ESTs will be 3'-end sequenced and incorporated into unigene sets for application to glass slide-based microarrays. We anticipate sequencing to be completed within the next six to nine months. We will begin alignment and assembly of our ESTs using either the StackPack or TIGR assembler software.

B.2.b. Expression analysis of *M. truncatula* mRNAs through the use of microarrays

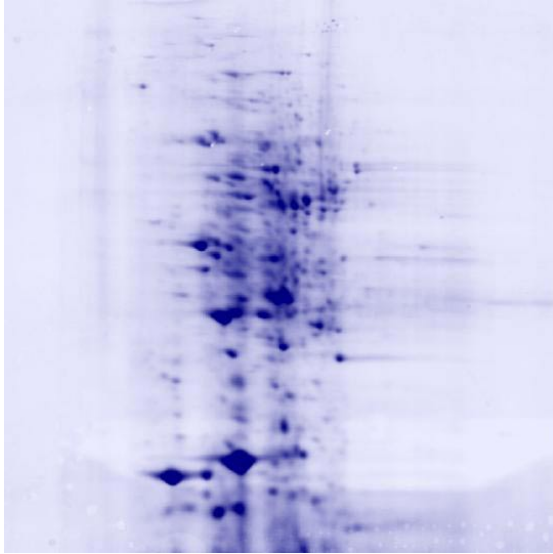
Changes in gene expression underlie many biological phenomena. The use of DNA microarrays will provide insights into elicitor- and UV-induced changes in gene expression on a global scale. cDNA microarrays are being generated using unique cDNA isolates identified in the *Medicago* EST program.

The infrastructure to generate, scan and analyze glass slide microarrays has been established at SRNF. Glass slide microarrays are being generated on the BioRobotics Ltd, *MicroGrid* system, hybridized with fluorescently labeled cDNA probes (Schena *et al.*, 1995; DeRisi *et al.*, 1997 and Kehoe *et al.*, 1999) and analyzed using the GSI Lumonics ScanArray 4000 two-color microarray analysis system.

Briefly, complex probes will be generated from mRNAs isolated from control and elicitor treated tissues using fluorescently labeled dNTPs in a first-strand cDNA synthesis. The signal ratio of differentially labeled (Cy3 vs. Cy5) fluorescent probe populations will be used to evaluate modulation in gene expression between control and treated samples, and treated samples over time (Schena *et al.*, 1995; DeRisi *et al.*, 1997). This core information resource will provide researchers with the capability to scan across multiple data sets in search of recurring patterns of expression (Kehoe *et al.*, 1999).

SRNF and the NSF-funded *M. truncatula* genome project have agreed to establish a standardized set of controls to be included on arrays generated from both programs. These experimental controls will increase the likelihood that data sets obtained from each program will be interchangeable. Standardized controls used in the Arabidopsis community will also be used. SRNF will also collaborate with the worldwide *Medicago truncatula* community in the generation of standardized "complete" *Medicago* microarray chips. Microarrays and their analysis facilities will be made available to the *Medicago* research community.

To supplement the expression analyses of data generated by microarrays, we will add an "open system" such as cDNA-AFLP or serial analysis of gene expression (SAGE) to our set of research tools. Such open system approaches allow for the identification and analysis of genes not previously characterized. With "closed systems" such as microarrays, analysis is limited to only those species previously identified and assigned to an array. Both cDNA-AFLP and SAGE have been used to study gene expression in plant systems (Durrant *et al.*, 2000; Matsumura *et al.*, 1999). Among the high-throughput, comprehensive technological methods used to analyze transcript expression levels, array-based hybridization and SAGE are currently the most common approaches. In a recent comparison of SAGE and array-based technologies (Ishii *et al.*, 2000) the two methods correlated quite well in both absolute expression analyses and comparative analyses during differentiation. The correlation was better for genes with higher expression levels and greater changes in expression. In the next 12 months, we will



determine which of the two above-mentioned open system technologies best suits our research needs. These efforts will be funded by SRNF.

B.2.c. Proteomic analysis

We propose to use two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) and mass spectrometry to investigate differences in protein expression between control and elicited cell cultures. 2D-PAGE has been established as the dominant technique for proteomic analysis (Blackstock *et al.*, 1999) since its introduction in 1975 (O'Farrell, 1975). The technique utilizes isoelectric focusing and polyacrylamide gel electrophoresis for first and second dimension separation, respectively. Currently, 2D-PAGE technology is capable of resolving some 10,000 proteins, with 2,000 proteins being routine (Klose *et*

al., 1995). A recent review describes the role of 2D-PAGE in proteomic and genetic studies of plant systems, including its use as a tool to investigate genetic diversity, phylogenetic relationships, mutant characterization, and drought tolerance (Thiellement *et al.*, 1999). 2D-PAGE has also been utilized to study plant defense-associated responses (Wagoner *et al.*, 1982), responses to MeJa (Mueller-Uri *et al.*, 1988), luminal and thylakoid proteins from chloroplasts (Peltier *et al.*, 2000), alterations in tomato plasma membrane profiles (Benabdellah *et al.*, 2000) and anoxia in maize roots (Chang *et al.*, 2000). Currently we are using 2D-PAGE to profile *M. truncatula* leaf, root, and membrane proteins. An example gel of a *M. truncatula* cytosolic root extract is provided in Fig. 2.

Although 2D-PAGE analysis has been used for the last 25 years in protein profiling, it provides limited information on protein identification. Recent advances in mass spectrometry and the establishment of protein databases have substantially increased the ease and speed with which proteins can be identified

Fig. 2. 2D-PAGE gel of *M. truncatula* root cytosolic proteins.

(Yates, 1998). The union of these technologies is the foundation for modern proteomic studies. We propose to use two MS techniques to identify proteins profiled by peptide mass-mapping of proteolytic digest fragments using MALDI-TOF-MS (Yates, 1998; Wolf *et al.*, 1998). The observed mass fragments can be searched against a theoretical list of proteolytic peptide masses predicted by a given database (Fig. 3). Increased peptide mass accuracy, capable with current MS instrumentation at SRNF, has been shown to increase the success and selectivity of such searches (Jensen *et al.*, 1996). Because amino acid sequence information is already available for nearly all the enzymes involved in flavonoid biosynthesis, and for early enzymes of saponin biosynthesis, the peptide mass mapping strategy should provide accurate identification of changes at the proteome level relating to the specific biochemical pathways targeted in this proposal.

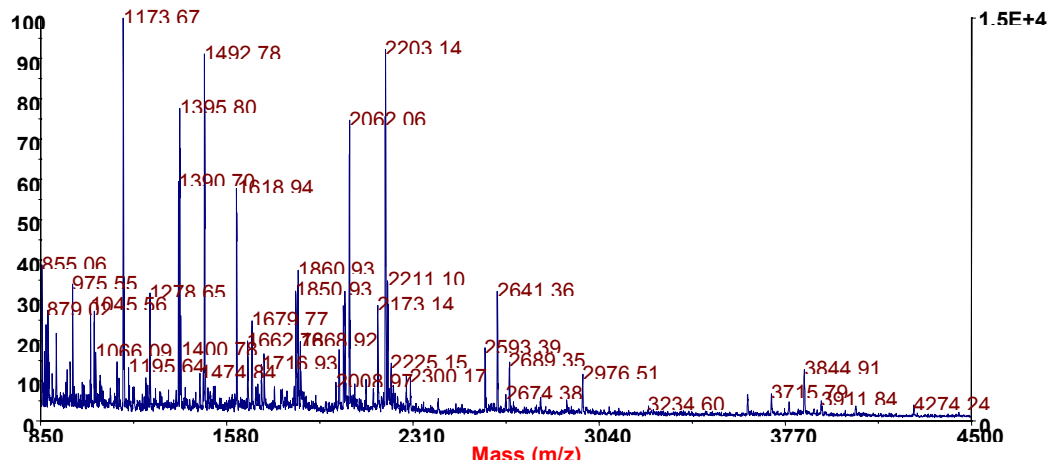


Fig. 3 – MALDI-TOF-MS peptide mass map of an excised spot from a *M. truncatula* 2D-PAGE gel that was digested in-gel using trypsin. The peptide mass map was queried against the protein databases and the protein identified as ATP synthase β chain.

The second technique is experimentally more complex but will be used if peptide mass map database queries are unsuccessful. This technique involves tandem mass spectrometry (MS/MS) to directly sequence peptides/proteins and to elucidate post-translational modifications (Yates, 1998). During the MS/MS experiment, only the peptide mass of interest is isolated or transmitted, thus discriminating against all other components of the mixture with different mass-to-charge values. After isolation, the peptide is further fragmented using a unimolecular or bimolecular (collision gas) strategy. The observed ion peaks in the tandem mass spectra can be used to elucidate a sequence or be submitted to a database using an alternate “sequence-tag” or alignment searching strategy. Sequence alignment procedures can yield protein identification and have the added benefit of simultaneous differentiation of sequence heterogeneity. We have validated this concept through successful determination of sequence heterogeneities using a developmental tandem TOF-TOF instrument (Collaboration with PE BioSystems, Framingham, MA) for in-gel digested samples of *M. truncatula* proteins (Wolf *et al.*, 2000). An example is illustrated in Fig. 4 for Rubisco from *M. truncatula*, the sequence of which is not in the protein database, but which could be identified by peptide mass mapping by comparison to alfalfa Rubisco large subunit. Most of the protein sequence was determined to be homologous to the alfalfa protein through observed peptides in the mass map. However, an unidentified peak at $m/z = 1019$ was observed and analyzed by tandem TOF-TOF-MS. Through sequence tag searching of the tandem TOF-TOF information we could identify this peptide as differing from the corresponding peptide from alfalfa by a change at residue seven from alanine to leucine.

The Noble Foundation currently possesses a MALDI-TOF instrument capable of high resolution and high mass accuracy peptide mass mapping. Unfortunately, this instrument is not suitable for high throughput tandem MS sequencing of peptides and the MALDI-TOF peptide mass mapping strategy is limited to the identification of proteins contained within the databases. We are therefore requesting matching funds for the purchase of a Q-TOF mass spectrometer to enhance our protein identification strategies using tandem MS, for novel protein sequencing, for elucidation of post-translation modifications, and for high through-put homology based sequencing of *M. truncatula* proteins (Wolf *et al.*, 2000). The Q-TOF instrument offers the greatest opportunities for successful sequencing of novel proteins and is the preferred choice because of its tandem MS range, mass accuracy and sensitivity. We

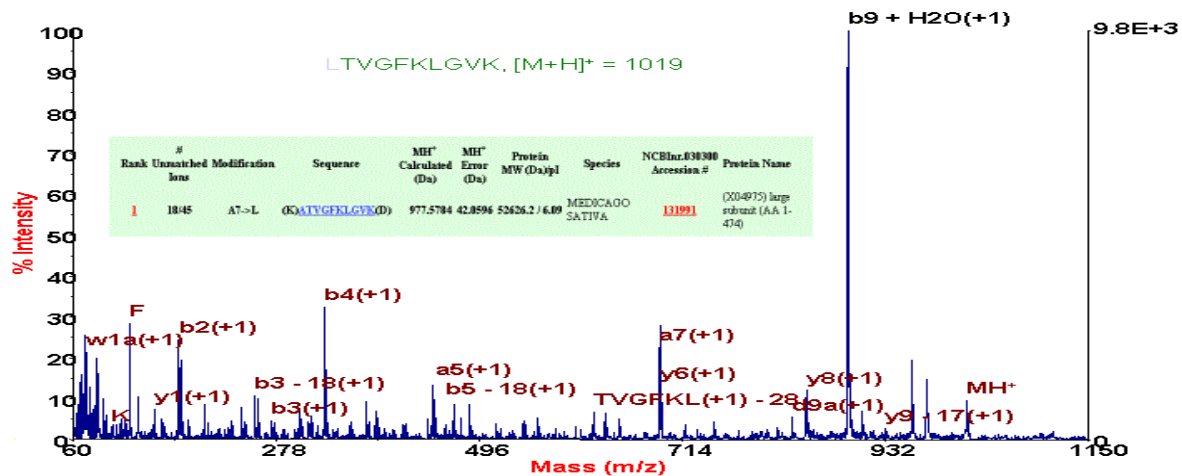


Fig. 4. Tandem TOF-TOF mass spectrum of an unidentified peptide observed in the peptide mass map of a *M. truncatula* protein. Database searching of this sequence reveals a single amino acid change from the database sequence of the corresponding peptide from alfalfa rubisco large subunit.

also request matching funds for the purchase of a capillary HPLC to provide tandem HPLC/MS/MS capabilities that will enable us to pursue alternate protein profiling and identification techniques. Capillary HPLC yields higher chromatographic resolution and significantly lowers the amount of material necessary for analysis. Although 2D-PAGE has the ability to profile large number of proteins, recent studies have shown that many proteins are not observed in 2D-PAGE experiments, for example proteins with low abundance or hydrophobic membrane proteins (Gygi *et al.*, 2000). Because of these limitations, alternative protein identification strategies are currently being developed including HPLC-MS/MS (Yates, 1998), isotope tagging (Gygi *et al.*, 1999) and accurate mass tags (Conrads *et al.*, 2000). The capillary HPLC/Q-TOF would allow us to pursue some of these developing technologies while we use our present instrumentation for the more established methods.

B.2.d. Metabolite profiling

Metabolite profiling provides a deeper insight into the ultimate functions of gene expression (Fiehn *et al.*, 2000) and is the key to understanding how changes at the level of the genome and proteome affect cellular function (Trethewey *et al.*, 1999; Glassbrook *et al.*, 2000). Unlike genomics and proteomics, a single analytical technique does not exist that is capable of profiling all the low molecular weight metabolites of the cell.

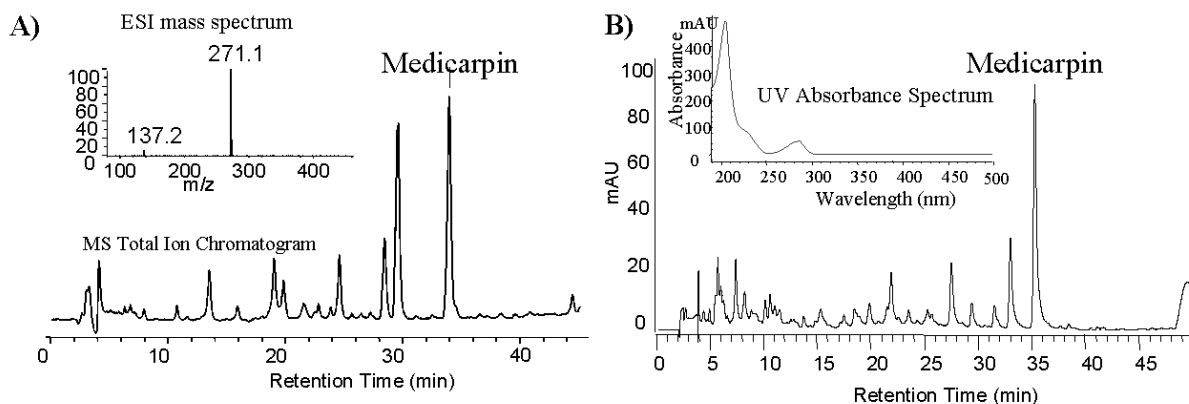


Fig. 5 - Profiling of the isoflavonoid medicarpin from a yeast elicitor-treated cell suspension culture of *M. truncatula*. Trace **A** shows a total ion chromatogram obtained using our Bruker Esquire three-dimensional quadrupolar ion-trap LC/MS. The inset shows the mass spectrum for the medicarpin peak, identical to that of authentic medicarpin. Trace **B** shows a similar HPLC separation, but with photodiode array detection. The inset shows the UV/visible absorption spectrum of the medicarpin peak.

We have been using traditional methods such as GC/MS to profile many metabolites including lignins (Jung *et al.*, 1993) and simple phenylpropanoids (Orr *et al.*, 1993a,b). Recently we have developed more sophisticated techniques revolving around HPLC/MS to profile many of the more difficult classes of metabolites including isoflavonoids (Fig. 5), related phenolic conjugates (Sumner *et al.*, 1996; Barnes *et al.*, 1998; Watson *et al.*, 1998) and saponins (Huhman *et al.*, 2000; Oleszek, 1988, Perez *et al.*, 1997; Oleszek *et al.*, 1990; Nowacka *et al.*, 1992). In spite of the development of GC/MS for profiling large numbers of compounds from *Arabidopsis* (Fiehn *et al.*, 2000), we believe that this method is not optimum for labile compounds or those for which chemical derivatization is not facile, and we therefore prefer a multifaceted approach. We will continue to expand the scope of our HPLC/MS methods and will seek to incorporate capillary electrophoresis (CE)/MS for enhanced profiling of soluble sugars, sugar phosphates, and complex carbohydrates (El Rassi, 1995). Our metabolite profiling approach will therefore use:

- HPLC/MS for flavonoids, isoflavonoids, phenylpropanoids and triterpene saponins
- HPLC with post-column derivatization and UV detection for amino acids
- CE/MS for carbohydrates
- GC/MS for terpenoids, lipids and lignins

An example of using HPLC/MS to profile and identify saponins in *M. truncatula* and alfalfa is provided below. Saponins are composed of triterpenoid or steroidal subunits that are substituted with a varying number of sugar side chains. The labile nature of the glycosidic bond prevents GC/MS profiling of these conjugates, and their analysis is further complicated by the lack of a strong chromophore, making them difficult to detect with traditional methods such as UV absorbance. The saponins from alfalfa and *M. truncatula* were separated and identified by reverse-phase HPLC and electrospray ionization (ESI) mass spectrometry (Fig. 6). Seventeen saponins were identified in alfalfa (cv. Apollo) based upon negative-ion ESI/LC/MS, ESI/LC/MS/MS and literature data. Negative-ion ESI/LC/MS and ESI/LC/MS/MS were utilized along with HPLC retention times to identify eighteen saponins in *M. truncatula* (cv. Jemalong). The saponin aglycone structures are shown in Fig. 6. We are not aware of any previous reports identifying saponin glycosides in *M. truncatula*.

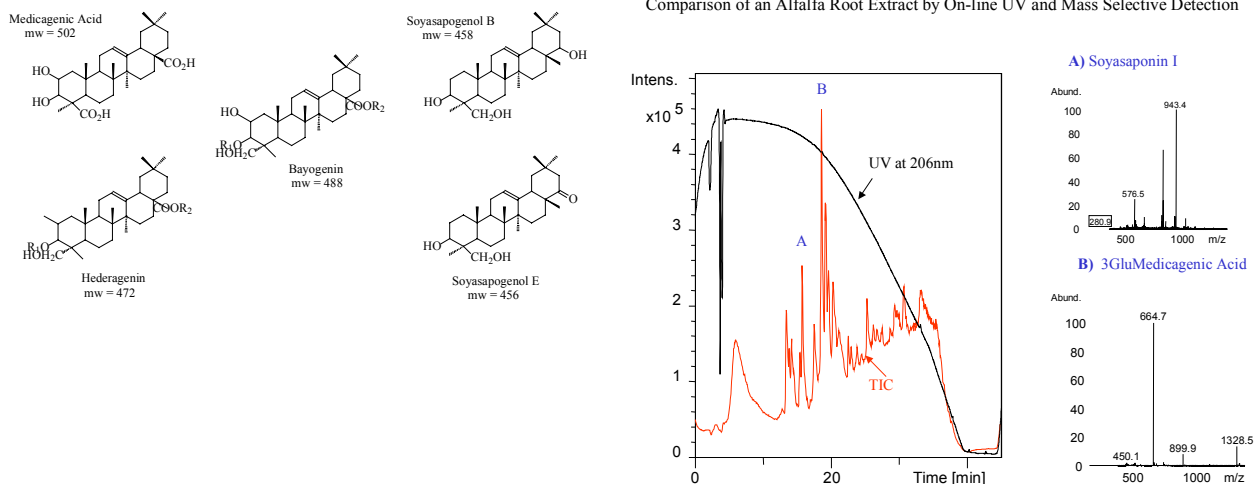


Fig. 6. Saponin structures and a comparison of UV vs. mass selective detection. MS detection is more sensitive than UV and provides selective chemical information in the form of molecular weight.

C. Informatics.

The purpose of collecting data from three levels of cellular function (mRNA, protein and metabolites) is to be able to characterize the state of the *Medicago truncatula* culture cells with a high level of detail. The data for each point in the time course of elicitation forms a "snapshot" of the cellular state, and the sequential presentation of data from all time points corresponds to a movie of how the cell responds to the external perturbations (Kell & Mendes, 2000; Mendes, 2001).

The analysis of the large volumes of data produced in our planned experiments is of primary importance. The value of such a rich data set lies in mining it for *i*) identification of the molecular function of genes, *ii*) identification of the molecular function of proteins, *iii*) relating protein spots on the 2D gel to ESTs in an array, *iv*) identification as to which genes and proteins are responsible for large changes in specific metabolite concentrations (e.g. to uncover pathways) and *v*) forming predictive models of the cellular responses to the elicitors used. Exploration of these rich data sets will require a variety of different analyses and computer simulations.

The characteristics of the research proposed here are quite challenging from a bioinformatics point of view. We will generate true functional genomics data, made up of three main data types: gene expression cDNA microarray data, 2D-PAGE proteomics data, and data from LC/MS and GC/MS metabolite profiles. The power of performing these analyses in parallel is to be able to make comparisons and associations between the different data sets. A further challenge is to make the system extensible in terms of the algorithms available to analyze the data. Currently there is no consensus as to what algorithms are best suited for this purpose, and indeed this is an area of intensive research and new methodologies are continually being suggested (e.g. Alter *et al.*, 2000). We will process the data with several different algorithms and compare the results as we proceed with this study. Given these requirements, we propose to construct a system comprised of a relational database, an analysis server, and client software for uploading and retrieving data. Fig. 7 depicts a data flow diagram including all the software components that we propose to build. This software architecture will allow the researcher to have a global view of all the data generated by this project. It will be made available to other academic researchers who can use it to manage data from similar projects.

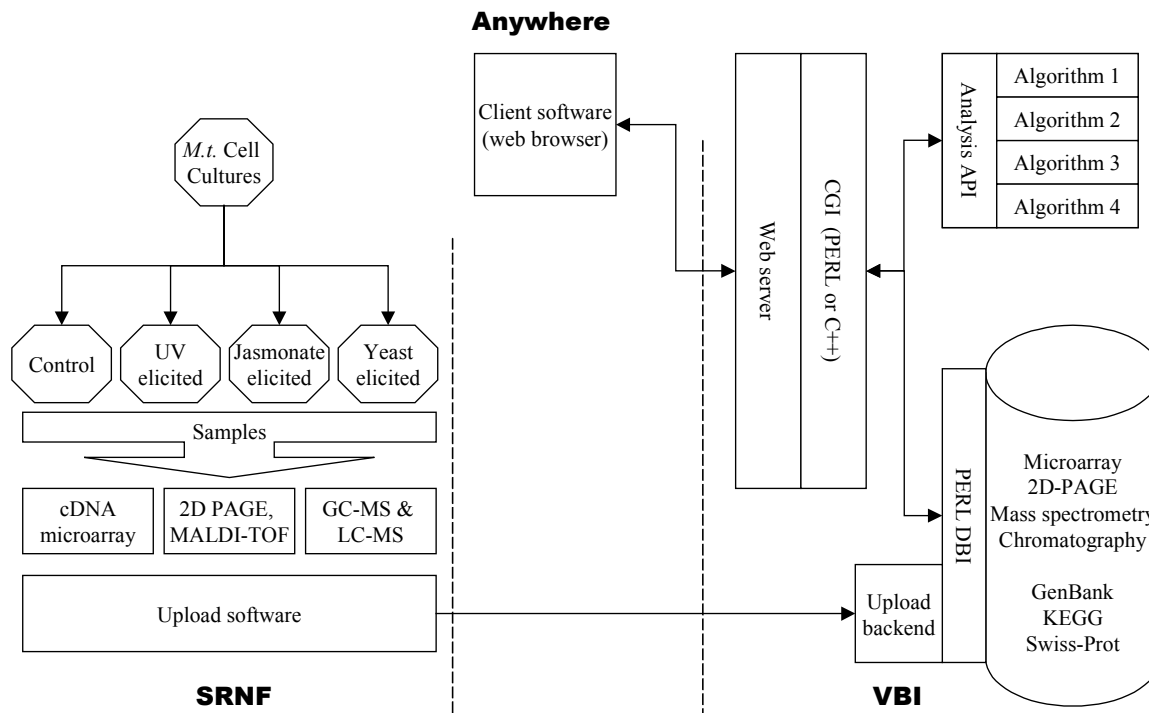


Figure 7 - Diagram representing the flow of data from the experiments to the database and from this to the analysis server and web-based client software. Dashed lines represent site boundaries (“Anywhere” means any browser with Internet connection).

C.1. Data storage

Because this project will generate data of several different types, it is important that the bioinformatics support system be capable of dealing with these in an integrated way. To achieve this goal we propose to store all the data in one relational database (Oracle 8). Such a database will have a large number of tables, needed to support all the disparate data we will collect. Some could argue that it would be best to store each major data type (microarray results, metabolic profiles, etc) in its own specific database. We have contemplated that alternative solution but this would require us to write or adopt an “integration layer” (usually called middleware by the software industry). Careful considerations based on our experience in such architectures (e.g. Siepel *et al.*, 2000) led us to conclude that for this specific project a single database will convey higher performance than competing designs. An added bonus is that this solution requires the shortest development time, thus allowing us to spend more time on researching data analysis algorithms rather than programming enterprise-type software.

For the cDNA microarray gene expression data we will construct a series of tables that will let us capture the intricate details of these experiments and their results. We will reuse, as much as possible, the data model defined for the Stanford Microarray Database (<http://genome-www4.stanford.edu/MicroArray/SMD/>). The images resulting from scanning the microarrays will be stored in files referenced in the appropriate tables of our database. We believe that it is important to supply other researchers with these unprocessed primary data, especially since we anticipate

that in the near future better methods to quantify mRNA levels from these images may appear and we want to allow third parties to analyze these raw data in a way that may be different from the one we used.

For 2D-PAGE/MALDI-MS/ESI-MS proteomic results we will construct appropriate tables in the database such that one can easily relate spots in the gel with other features of our experiments. Each spot will be related to its coordinates on the gel, its approximate molecular weight, its approximate *pI*, its area, its density, its mass spectrum (when it has been recorded), and indeed with the protein identity when determined. As with microarrays, the actual scanned images will be stored as files but will be referenced in the relational database, allowing anyone to retrieve the original image. Comparative image analysis of 2D-PAGE gels will be carried out by commercial software provided by Nonlinear, PLC.

Metabolite profiling data consist of chromatograms from various instrument platforms that associate mass or electromagnetic absorption spectra with each chromatographic peak. We will create tables for each type of spectrum/chromatogram; in this case these are the raw data and will reside in the database itself so it can be queried directly. Ideally all peaks in the chromatograms would be identified as compounds and their concentrations quantified. Our schema will allow any peak of a chromatogram to be annotated with the compound identification as the compounds comprising the peaks become known. These annotations will provide the link to the proteomic and gene expression databases since the compounds are derived from enzymes, and hence from genes, through metabolic pathways.

Interrelating the above three data types passes through use of reference databases of nucleotide and protein sequences and metabolic pathways. Many of the existing resources have resulted from previous public funding (either US or from foreign governments). We will use GenBank for nucleotide sequences, Swiss-Prot/Trembl for protein sequences, and KEGG for metabolic pathways. Realizing that the latter has a very poor coverage of legumes and none of *Medicago truncatula*, we will supplement our own local copy of KEGG with information from legume metabolic pathways recovered from the literature and from SoyBase. We will make these pathway data available to the public and indeed to the main KEGG repository and other public pathway databases.

C.2. Analyses server

An important part of this project will consist of research to determine which analysis methods are best suited to establish informative relationships among all of the data. To allow maximum flexibility and creativity for researchers in the exploratory phase of this investigation we propose to construct an application server that will have: *i*) logic to retrieve the desired data from the database, *ii*) a well defined protocol by which any algorithm required by an investigator can be plugged into this server, and *iii*) a mechanism to present a choice of algorithms that can be used to process specific data selected by the user (retrieved from the database). This architecture permits us to extend the analysis server to allow the use of any algorithm at any time, and still provides a single point of entry where the scientists from any institution can come to access and analyze the data. The extensibility of this component is crucial to the reality of this project in which some researchers will be analyzing data while, in parallel, others will be developing new methods of analysis. By implementing all of the analysis algorithms on a server there will be no need to pass the mass of raw data through to the client software, and this will allow us to implement any heavy computational algorithm on high performance computers at VBI.

A major focus of this project is to research many algorithms for their usefulness in processing the data to answer the biological questions enumerated in the introduction. We will apply those algorithms that are already in use by other groups (clustering, self-organized maps, etc.) but also investigate a number of others. Table 2 describes some algorithms that we will research first. The list is by no means comprehensive, as we do not want to bias ourselves towards *any* algorithm before comparison of their performances. The analysis server will act as a selectable menu of data processing algorithms that researchers will have available.

Table 2 - Some classes of algorithms that will be applied for analysis of the experimental results of this project. Special emphasis will be put on correlating information from gene expression with proteomics and metabolism

Analysis Methods	Description	References
Clustering	Group ESTs, proteins and metabolites by correlation of their time courses.	Fayyad <i>et al.</i> (1998) Kaufman & Rousseeuw (1990) Sokal & Michener (1958)
Principal component analysis (PCA), Projection pursuit regression (PPR), Singular value decomposition (SVD)	Find a small number of combinations of mRNAs, proteins and metabolites that explain most of the observed behaviour	Hotelling (1933) Jones & Sibson (1987) Alter <i>et al.</i> (2000)
Classifier systems	Construct systems, such as decision trees, to classify ESTs, proteins and metabolites according to their time courses	James (1985) Quinlan (1993)
Time series analysis	Determine frequencies in the time series, and possibly smooth the time courses. Determine the dimension of the time series.	Brockwell & Davis (1996)
Metabolic Control Analysis	Estimate metabolic control coefficients and co-control coefficients (quantitative measures of rate limitation).	Fell (1996)
System identification / Parameter estimation	Find kinetic models that best suit the observed time courses for ESTs, proteins and metabolites.	Ljung & Ljung (1998) Mendes & Kell (1998)
Continuous-time recurrent neural networks	Reverse engineering of genetic networks from time course data	D'Haeseleer <i>et al.</i> (1999) Wahde & Hertz (2000)

C.3. Upload software

A great source of reluctance among researchers in actually submitting their data to databases is that problems frequently arise when formatting the data appropriately. This is especially a concern when the data producers are not the same people who manage the database. To minimize this problem, we will construct, in the first year, an upload tool to be deployed at SRNF where the data are being generated. The upload software will verify that the data have met all of the schema requirements before allowing them to be entered in the database. It will also maintain a local database of protocols and experimental procedures so that the SRNF researchers need only enter the descriptions of the technical and experimental condition protocols once. Thereafter they simply select the appropriate protocol from a list. This software will also enforce a well-defined nomenclature for classifying experiments. Without a controlled vocabulary, queries based on experimental conditions leading to complex data analyses would quickly become unmanageable. By implementing and using this upload software from the beginning, we will minimize both time required to format and submit the data and the errors associated with independent typing events.

Given the large volume of data to be generated, the mechanism for upload to the database proper will not be through the Internet. Instead, data will be sent by Federal Express or UPS from SRNF to VBI on magnetic tapes or optical disks. The upload software will be the means to write data to the media, and this will be done in such a way that they will be ready to be imported to the appropriate databases via

simple SQL scripts. The data will also be stored at SRNF for redundancy, but only on copies of the physical transfer media. The on-line data storage medium at VBI, however, will be a fault-tolerant disk array (see below).

C.4. Data retrieval, analysis and visualization software

Above we described a mechanism by which all of the analyses are carried out on the server side, including the pre-processing for data visualization. This implies that the client software will not be very sophisticated. We will implement the client software based on the world-wide-web. It will be composed of HTML pages that will load onto the researchers' web browsers and that will connect to CGI scripts on the VBI web server. These scripts will then process the requests via database queries and/or the calls to the analysis server, depending on whether the request is to visualize data or to analyze them. They will subsequently format the material to send to the client (raw data or analysis results) through HTML and graphics. This mechanism also permits us to add links to relevant external sites and to provide a portal to our data to external sites easily, without requiring specialized software. We believe that this mechanism is the most flexible, easy to implement and the most familiar to the research community. It is also a proven method that is used in numerous databases such as GenBank, SwissProt, KEGG, and others.

C.5. Integrative modeling and computer simulation – A systems approach to the biology of elicitation in *M. truncatula*.

Scientific knowledge progresses through hypothesis forming and testing. On the one hand, the experiments planned for this project are designed to uncover relations between levels of mRNA, proteins and metabolites with the view to discovering the function of novel genes. This is mostly a data mining activity, or *analysis*, by which we will use the measurements of the variables (the data) to infer the values of the parameters, *i.e.* the invariable properties of the biological system. This is known in mathematics as an *inverse problem* (Mendes & Kell 1996, 1998). On the other hand, novel whole-cell approaches such as genomics can only be used to their full potential if *synthesis* is also involved. In this case one uses those parameters determined in the analyses to reconstruct the essential properties of the system (a direct problem, reflecting what happens in nature: the parameters determine the variables). This, in practice, will be carried out by computer simulation of the model formed by the analyses and by the underlying assumptions. By comparing the behavior of the simulations with the outcome of the experiments we will either *i)* gain confidence that the analytic results plus the assumptions are indeed consistent with the observations or *ii)* prove that the underlying assumptions are not correct and a new model needs to be constructed, iterating the process as many times as necessary.

The PI has extensive experience in this field and is the author of one of the most popular software packages for simulation of biochemical systems, Gepasi (Mendes 1993, 1997, see also <http://www.gepasi.org>). Gepasi has been freely available and supported since 1990. We will use this software to simulate the complete time-courses of elicitation (*i.e.* metabolites, protein and mRNA levels) in years 3 and 4. One of the outcomes of the computer simulations will be, for example, that we demonstrate the requirement for a particular gene in the chain of events that leads to an elicited cellular response. With computer simulation we can answer "what-if" questions and thus see the consequences of suppressing certain genes and their gene products from the system – thereby suggesting their roles in specific pathways. We believe that without this step the results of the analyses would be without validation and that this unique data set would not be used to its full potential. Simulation should not be seen as a substitute to experiments. It is rather a rigorous way of formulating hypotheses that will lead to new experimentation. For large dynamical systems we argue that this is the most efficient way in which one can proceed in the scientific process, because in such systems the dynamics are often counter-intuitive to the human mind. This is the main subject of an emerging field which is being referred to as "systems biology". Last November the PI participated in the program committee and as a speaker (Mendes, 2001) on the 1st International Conference on Systems Biology in Tokyo.

C.6. Volume of data

The project proposed here would generate a substantial amount of data. Table 3 summarizes our estimates for the total computer storage requirements. These data will be archived in a relational database system described above which will reside in a Sun Microsystems E450 server equipped with 200Gb (usable) RAID 5 system.

Table 3 - Predicted volume of data to be produced in this project.

Data type	Description	Size/sample	Samples	Total size
Gene expression	1 High res. color bitmap image	15 Mb	252	3,276 Mb
Proteomics	1 High res. grayscale bitmap image + 100 MALDI-TOF/MS	103 Mb	252	25,956 Mb
Sugars profile	1 Capillary electrophoresis/MS chromatogram	50 Mb	279	13,950 Mb
Amino acids profile	1 HPLC chromatogram	2.5 Mb	279	698 Mb
Lipids profile	1 GC/MS chromatogram	2 Mb	279	558 Mb
Flavonoids / isoflavonoids profile	1 LC/MS chromatogram	50 Mb	279	13,950 Mb
Phenylpropanoids profile	1 LC/MS chromatogram	50 Mb	279	13,950 Mb
Terpenes / saponins profile	1 LC/MS and 1 GC/MS chromatogram	52 Mb	279	14,508 Mb
Total				130,726 Mb

C.7. Preexisting software

In order to construct the bioinformatics support system for this project we will re-use a series of preexisting software. Our intention is to re-use as much software as possible that is free to academics such that our system can later be re-used by others. The extant software packages that we will re-use are:

i) Gepasi, a metabolic pathways simulator. This has been written by the PI throughout the last 10 years and has received past funding from the Portuguese JNICT, and the British BBSRC. Gepasi is widely distributed for no fee and has a well-established user base (see <http://www.gepasi.org/gep3res.html> for a list of research publications that used Gepasi). This software simulates time courses of biochemical pathways given the kinetics of their enzymes. It can also be applied to hierarchical systems where transcription and translation is included. The program also fits kinetic models of full pathways (as opposed to isolated enzymes) to observed time courses. It will be used in this project in the latter capacity as well as in constructing an integrative model of the elicitation time courses.

ii) Oracle database server. This is mostly free to academics in the US for purposes of scientific research. We will also make sure that the database scripts use only the subset of SQL that is common to the free relational database management systems such as PostgreSQL. That will ensure that anyone can use our software without paying large license fees.

iii) Numerical routines archived in the Netlib (<http://www.netlib.org>) and Statlib (<http://lib.stat.cmu.edu>) servers. These are public domain algorithms that are well optimized and debugged

and are widely used in scientific computing. In particular the codes for clustering, PCA, and SVD (see Table 1), that will be used in our analysis server will come from these sources.

iv) OpenDX, an open source visualization package (<http://www.opendx.org>). This is a popular and powerful visualization package previously developed and commercialized by IBM but which is now an open source project. This will be used for producing specific views of the data.

v) XGobi and XGvis (Swayne *et al.*, 1998), two visualization software packages freely available (<http://www.research.att.com/stat/xgobi/>). These will be mostly used for producing interactive dynamic graphics, such as projection pursuit and grand tours.

vi) Phoretix 2D-PAGE analysis software by Nonlinear Dynamics, Ltd. (<http://www.nonlinear.com/>). This commercial software will be used for the comparative analysis of 2D-PAGE gels. It performs automated and manual spot detection, spot quantification, annotation, web page construction, and data export.

vii) Mass Transit by Palisade (<http://64.80.33.167/>) is translator of GC/MS data (chromatogram/mass and library) formats from Hewlett-Packard, Finnigan, Perkin-Elmer, Shimadzu, Varian and many more. This software will be used to convert mass-spectral data from our various instruments to a uniform file format useable by our database system.

viii) Grams by Galactic Industries (<http://www.galactic.com/>) is a comprehensive data processing and data management tool that can automatically recognize and read data files from hundreds of analytical instruments including NMR, FT-IR, Raman, NIR, GC-MS, UV-Vis, fluorescence, and chromatography instrumentation. This software will be used to import, convert, and export proprietary datafile formats to a uniform file format to be used by our database system.

D. Roles of participants

Pedro Mendes. Principal investigator. Overall project co-ordination. In charge of the development of proteomic and metabolic tables of the database, HTML-CGI software and analysis server. Supervision of programmers and one postdoc involved in constructing quantitative integrative cellular models of the whole biological system in years 3 and 4 and co-supervise a postdoc involved in the investigation of appropriate statistical and machine learning algorithms for analyzing the data.

Richard Dixon. Co-Principal Investigator. Coordination of the efforts at SRNF (generation of the biological data). Supervision of one technician for production and elicitation of cell cultures and analytical sample work-up (years 1 and 2) and two postdocs involved in metabolite analysis and mining of data sets in years 2-4.

Lloyd Sumner. Collaborator. Supervision of proteomic and metabolic profiling analysis in the SRNF Biological Mass Spectrometry Laboratory. Coordination of the metabolite profiling effort at SOSU. Supervision of two postdocs and two research assistants (laboratory technicians) *i.e.* two FTEs for proteomic analysis, two FTEs for development and implementation of metabolic profiling (years 1-3).

Greg May. Collaborator. In charge of microarray analysis at SRNF. Supervision of one postdoc for mining of microarray data in years 3 and 4.

Jennifer Weller. Collaborator. In charge of microarray profile analysis and development of corresponding tables in the database. Will develop and coordinate the workshop described in the Training and Outreach section. Will co-supervise one post-doc during years 2-4, involved in the investigation of appropriate algorithms and modeling techniques for analyzing the data.

Tim Smith. Collaborator. Sub-contract for metabolic profiling of carbohydrates and amino acids at SOSU (years 2 and 3). Supervision of two undergraduate students in years 2 and 3.

E. Training and Diversity

VBI and SRNF propose to institute a short course in "Bioinformatics and Functional Genomics of Legumes" under the auspices of this project, to be held for two weeks each summer of the granting period. The course is envisaged as encompassing both laboratory and bioinformatics components. SRNF has the conference center facilities to house students, lecture facilities for teaching courses and sufficient room to provide laboratory space for students. SRNF does not seek funding for the usage of the facilities in these courses. There will be a broad introduction into current issues in legume functional genomics and general areas of interest and research in bioinformatics, followed by specific questions for students to investigate. These will be developed by the staff to ensure that meaningful results can be attained in the allotted time and with the technologies available, or alternatively proposals may be generated by applicants to the course and be part of the selection process. Feedback will be actively sought from the course participants year to year to determine how the course can best evolve in this rapidly changing field. In the first year the course will be given to internal staff from the two institutions and a small number of collaborators of the senior staff in this project. Their feedback will be used to improve the course to be offered generally in years two, three, and four. This NSF-sponsored course will be advertised at the Plant and Animal Genome meetings, in journals such as *Science*, newsletters such as the *Agricultural Genomics Newsletter*, and in the Internet in appropriate usenet newsgroups (such as *bionet.jobs.offered*, *bionet.genome.arabidopsis* and *bionet.plants*). Students will be selected from among the applicants with a view to broadening the spectrum of researchers with both skill sets and including criteria of equal opportunity for such training.

Post-doctoral and Graduate research assistants and students to be hired in this project will be selected on the basis of equal opportunities. VBI, as part of Virginia Polytechnic Institute and State University, follows the latter's equal opportunity's and affirmative action procedures (see <http://neelix-fbox.cc.vt.edu/admin/eoaa/>).

Post-doctoral researchers from each institution funded by this project will spend time at each other's laboratories so that they can learn all the aspects of this multidisciplinary project. We anticipate that these mutual visits could take up to 6 months and will be decided on a case-by-case basis. A major consideration will be for each individual to become familiar with at least one other technique which is outside their own area of expertise.

D. Literature cited

- Alter, O., Brown, P.O. and Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences USA* 97, 10101-10106 (2000)
- Auh, C.K. and Murphy, T.M. Plasma membrane redox enzyme is involved in the synthesis of O₂⁻ and H₂O₂ by *Phytophthora* elicitor-stimulated rose cells. *Plant Physiology* 107, 1241-1247 (1995)
- Ballance, G.M. and Dixon, R.A. *Medicago sativa* cDNAs encoding chalcone reductase. *Plant Physiology* 107, 1027-1028 (1994)
- Batz, O., Logemann, E., Reinold, S. and Hahlbrock, K. Extensive reprogramming of primary and secondary metabolism by fungal elicitor or infection in parsley cells. *Biological Chemistry* 379, 1127-1135 (1998)
- Barnes, K. A., Smith, R. A., Williams, K., Damant, A. P. and Shepherd, M. J. A microbore high performance liquid chromatography/electrospray ionization mass spectrometry method for the determination of the phytoestrogens genistein and daidzein in comminuted baby foods and soya flour. *Rapid Communications in Mass Spectrometry* 12, 130-138 (1998)
- Barz, W. and Mackenbrock, U. Constitutive and elicitation induced metabolism of isoflavones and pterocarpan in chickpea (*Cicer arietinum*) cell suspension cultures. *Plant Cell, Tissue and Organ Culture* 38, 199-211 (1994)
- Bell, C., Dixon, R.A., Farmer, A.D., Flores, R., Inman, J., Gonzales, R.A., Harrison, M.J., Paiva, N.L., Scott, A.D., Weller, J.W. and May, G.D. The *Medicago* genome initiative: a model legume database. *Nucleic Acids Research* 29, 114-117 (2001)

- Benabdellah, K., Azcón-Aguilar, C. and Ferrol, N., Alterations in the plasma membrane polypeptide pattern of tomato roots (*Lycopersicon esculentum*) during the development of arbuscular mycorrhiza. *Journal of Experimental Botany* 51, 747-754 (2000).
- Blackstock, W.P. and Weir, M.P. Proteomics: quantitative and physical mapping of cellular proteins. *Trends in Biotechnology* 17, 121-127 (1999)
- Brockwell P.J. and Davis R.A. *Introduction to time series and forecasting*. New York: Springer-Verlag (1996)
- Broekaert, W.F., Terras, F.R.G., Cammue, B.P.A. and Osborn, R.W. Plant defensins: Novel antimicrobial peptides as components of the host defense system. *Plant Physiology* 108, 1353-1358 (1995)
- Carrier, D.J., Cosentino, G., Neufeld, R., Rho, D., Weber, M. and Archambault, J. Nutritional and Hormonal requirements of Ginkgo biloba embryo-derived callus and suspension cell culture. *Plant Cell Reports* 8, 635-638 (1990)
- Chang, W.W.P., Huang, L., Shen, M., Webster, C., Burlingame, A.L. and Roberts, J.K.M., Patterns of protein synthesis and tolerance of anoxia in root tips of maize seedlings acclimated to a low-oxygen environment, and identification of proteins by mass spectrometry. *Plant Physiology*, 122, 295-317 (2000).
- Chappell, J. and Hahlbrock, K. Transcription of plant defence genes in response to UV light or fungal elicitor. *Nature* 311, 76-78 (1984)
- Christie, W.W. *Lipid Analysis*. 2nd Edition, Pergamon Press, Elmsford, New York (1982)
- Conconi, A., Smerdon, M.J., Howe, G. A. and Ryan, C. A. The octadecanoid signaling pathway in plants mediates a response to ultraviolet radiation. *Nature* 383, 826-829 (1996)
- Conrads, T. P., Anderson, G.A., Veenstra, T.D., Paša-Tolic, L. and Smith, R.A., Utility of accurate mass tags for proteome-wide protein identification. *Analytical Chemistry* 72, 3349-3354 (2000)
- Cook, D.R. *Medicago truncatula* - a model in the making! *Current Opinion in Plant Biology* 2, 301-304 (1999)
- Dangl, J.L., Hauffe, K.D., Lipphardt, S., Hahlbrock, K. and Scheel, D. Parsley protoplasts retain differential responsiveness to U.V. light and fungal elicitor. *EMBO Journal* 6, 2551-2556 (1987)
- DeRisi, J.L., Iyer, V.R. and Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680-686 (1997)
- Desikan, R., Hancock, J.T., Coffey, M.J. and Neill, S.J. Generation of active oxygen in elicited cells of *Arabidopsis thaliana* is mediated by a NADPH oxidase-like enzyme. *FEBS Letters* 382, 213-217 (1996)
- Dixon, R.A. Plant tissue culture methods in the study of phytoalexin induction. In *Tissue Culture Methods for Plant Pathologists* (Ingram, D.S. and Helgeson, J.P., eds), pp. 185-196, Blackwell Scientific Publications, Oxford (1980)
- Dixon, R.A. Isoflavonoids: biochemistry, molecular biology, and biological functions. in *Comprehensive Natural Products Chemistry* (Vol. 1) (Sankawa, U., ed.), pp. 773-823, Elsevier, Oxford (1999)
- Dixon, R.A., Browne, T. and Ward, M. Modulation of L-phenylalanine ammonia-lyase by pathway intermediates in cell suspension cultures of dwarf French bean (*Phaseolus vulgaris* L.). *Planta* 150, 279-28 (1980)
- Dixon, R.A., Dey, P.M., Murphy, D.L. and Whitehead, I.M. Dose responses for *Colletotrichum lindemuthianum* elicitor-mediated enzyme induction in French bean cell suspension cultures. *Planta* 151, 272-280 (1981)
- Dixon, R.A., Harrison, M.J. and Paiva, N.L. The isoflavonoid phytoalexin pathway: from enzymes to genes to transcription factors. *Physiologia Plantarum* 93, 385-392 (1995)
- Doares, S.H., Narváez-Vásquez, J., Conconi, A. and Ryan, C.A. Salicylic acid inhibits synthesis of proteinase inhibitors in tomato leaves induced by systemin and jasmonic acid. *Plant Physiology* 108, 1741-1746 (1995)
- Donahue, B.A., Yin, S., Taylor, J.S., Reines, D. and Hannawalt, P.C. Transcript cleavage by RNA polymerase II arrested by a cyclobutane pyrimidine dimer in the DNA template. *Proceedings of the National Academy of Sciences USA* 91, 8502-8506 (1994)
- D'Haeseleer, P., Wen, X., Fuhrman, S. and Somogyi, R. Linear modeling of mRNA expression levels during CNS development and injury. *Pacific Symposium of Biocomputing* 4, 41-52 (1999)
- Durrant, W.E., Rowland, O., Piedras, P., Hammond-Kosack, K.E. and Jones, J.D.G. cDNA-AFLP Reveals a striking overlap in race-specific resistance and wound response gene expression profiles. *Plant Cell* 12, 963-977 (2000)
- El Rassi, Z. and Smith, J.T. Other direct and indirect detection methods of carbohydrates in HPLC and HPCE. In *Carbohydrate Analysis: High Performance Liquid Chromatography and Capillary Electrophoresis*, Z. (ed. El Rassi), Journal of Chromatography Library-volume 58, Elsevier, Amsterdam (1995)
- Fahrendorf, T. and Dixon, R.A. Stress responses in alfalfa XVIII: Molecular cloning of the elicitor-inducible cinnamic acid 4-hydroxylase cytochrome P450 from alfalfa. *Archives of Biochemistry and Biophysics* 305, 509-515 (1993)

- Fahrendorf, T., Ni, W., Shorrosh, B.S. and Dixon, R.A. Stress responses in alfalfa (*Medicago sativa* L.) XIX. Transcriptional activation of oxidative pentose phosphate pathway genes at the onset of the isoflavonoid phytoalexin response. *Plant Molecular Biology* 28, 885-900 (1995)
- Fayyad, U.M., Reina, C. and Bradley, P.S. Initialization of iterative refinement clustering algorithms. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* New York, Kluwer Academic Publishers (1998)
- Fell, D.A. *Understanding the Control of Metabolism*. Portland Press, London (1996)
- Felton, G.W., Korth, K.L., Bi, J.L., Wesley, S.V., Huhman, D.V., Mathews, M.C., Murphy, J.B., Lamb, C. and Dixon, R.A. Inverse relationship between systemic resistance of plants to microorganisms and to insect herbivory. *Current Biology* 9, 317-320 (1999)
- Fiehn, O., Kopka, J., Dörmann, P., Altmann, T., Trethewey, R.N. and Willmitzer, L. Metabolic profiling for plant functional genomics. *Nature Biotechnology* 18, 1157-1161 (2000).
- Franklin, C.I. and Dixon, R.A. Initiation and maintenance of callus and cell suspension cultures. In *Plant Cell Culture: A Practical Approach*. 2nd Edition (Dixon, R.A. and Gonzales, R.A., eds), pp. 1-25, Oxford University Press (1994)
- Funk, C., Gugler, K. and Brodelius, P. Increased secondary product formation in plant cell suspension cultures after treatment with a yeast carbohydrate preparation (elicitor). *Phytochemistry* 26, 401-405 (1987)
- Gerrish, C., Robbins, M.P. and Dixon, R.A. Cinnamic acid as a modulator of chalcone isomerase in bean (*Phaseolus vulgaris*) cell suspension cultures. *Plant Science Letters* 38, 23-27 (1985)
- Glassbrook, N., Beecher, C. and Ryals, J., Metabolic profiling on the right path. *Nature Biotechnology* 18, 1142-1143 (2000)
- Gowri, G., Paiva, N.L. and Dixon, R.A. Stress responses in alfalfa (*Medicago sativa* L.) XII. Sequence analysis of phenylalanine ammonia-lyase (PAL) cDNA clones and appearance of PAL transcripts in elicitor-treated cell cultures and developing plants. *Plant Molecular Biology* 17, 415-429 (1991)
- Grosskopf, D.G., Felix, G. and Boller, T. K-252a inhibits the response of tomato cells to fungal elicitors *in vivo* and their microsomal protein kinase *in vitro*. *FEBS Letters* 275, 177-180 (1990)
- Gundlach, H., Müller, M.J., Kutschan, T.M. and Zenk, M.H. Jasmonic acid is a signal transducer in elicitor-induced plant cell cultures. *Proceedings of the National Academy of Sciences USA* 89, 2389-2393 (1992)
- Guo, L. and Paiva, N.L. Molecular cloning and expression of alfalfa (*Medicago sativa* L.) vestitone reductase, the penultimate enzyme in medicarpin biosynthesis. *Archives of Biochemistry and Biophysics* 320, 353-360 (1995)
- Gygi, S.P., Corthals, G.L., Zhang, Y., Rochon, Y. and Aebersold, R., Evaluation of two-dimensional electrophoresis-based proteome analysis technology. *Proceedings of the National Academy of Sciences USA* 97, 9390-9395 (2000)
- Gygi, S. P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H. and Aebersold, R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology* 17, 994-999 (1999)
- Hahlbrock, K. and Scheel, D. Physiology and molecular biology of phenylpropanoid metabolism. *Annual Review of Plant Physiology* 40, 347-369 (1989)
- Harrison, M.J. and Dixon, R.A. Isoflavonoid accumulation and expression of defense gene transcripts during the establishment of vesicular arbuscular mycorrhizal associations in roots of *Medicago truncatula*. *Molecular Plant-Microbe Interactions* 6, 643-654 (1993)
- He, X.-Z., Reddy, J.T. and Dixon, R.A. Stress responses in alfalfa (*Medicago sativa* L.) XXII. cDNA cloning and characterization of an elicitor-inducible isoflavone 7-O-methyltransferase. *Plant Molecular Biology* 36, 43-54 (1998)
- Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24, 417-441, 498-520. (1933)
- Huhman, D.H., Dixon, R.A. and Sumner, L.W., Profiling saponin glycosides in *Medicago sativa* (alfalfa) and *Medicago truncatula* using HPLC coupled to an electrospray ion-trap mass spectrometer. *Proceedings of the 46th ASMS Conference on Mass Spectrometry and Allied Topics*, Palm Springs, CA (2000)
- Ishii, M., Hashimoto, Si., Tsutsumi, S., Wada, Y., Matsushima, K., Kodama, T. and Aburatani, H. Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis. *Genomics* 68, 136-143 (2000)
- James, M. *Classification algorithms*. New York: Wiley (1985)

- Jensen, O. N., Podtelejnikov, A. and Matthias Mann, M. Delayed extraction improves specificity in database searches by matrix-assisted laser desorption/ionization peptide maps. *Rapid Communications in Mass Spectrometry* 10, 1371-1378 (1996)
- Jones, M.C. and Sibson, R. What is projection pursuit? *Journal of the Royal Statistical Society A* 150, 1-36 (1987)
- Jung, H. G., Buxton, D. R., Hatfield, R. D. and Ralph, J. (Eds) *Forage Cell Wall Structure and Digestibility*. American Society of Agronomy, Crop Science of America, Soil Science Society of America (1993)
- Junghans, H., Dalkin, K. and Dixon, R.A. Stress responses in alfalfa (*Medicago sativa* L.) XV. Characterization and expression patterns of members of a subset of the chalcone synthase multigene family. *Plant Molecular Biology* 22, 239-253 (1993)
- Kaufman, L. and Rousseeuw, P.J. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York, Wiley (1990)
- Kawalleck, P., Plesch, G., Hahlbrock, K. and Somssich, I.E. Induction by fungal elicitor of S-adenosyl-L-methionine synthetase and S-adenosyl-L-homocysteine hydrolase mRNAs in cultured cells and leaves of *Petroselinum crispum*. *Proceedings of the National Academy of Sciences USA* 89, 4713-4717 (1992)
- Kehoe, D.M., Villand, P. and Somerville, S. DNA microarrays for studies of higher plants and other photosynthetic organisms. *Trends in Plant Science* 4, 38-41 (1999)
- Kell, D.B. and Mendes, P. Snapshots of systems: metabolic control analysis and biotechnology in the post-genomic era. In *Technological and Medical Implications of Metabolic Control Analysis* (eds. A. Cornish-Bowden and M. L. Cardenas), Kluwer Academic Publishers, Dordrecht, pp. 3-25 (2000)
- Kessmann, H., Edwards, R., Geno, P. and Dixon, R.A. Stress responses in alfalfa (*Medicago sativa* L.) V. Constitutive and elicitor-induced accumulation of isoflavonoid conjugates in cell suspension cultures. *Plant Physiology* 94, 227-232 (1990)
- Klose, J. and Kobalz, U. Two-dimensional electrophoresis of proteins: An updated protocol and implications for a functional analysis of the genome. *Electrophoresis* 16, 1034-1059 (1995)
- Kombrink, E. and Hahlbrock, K. Dependence of the level of phytoalexin and enzyme induction by fungal elicitor on the growth stage of *Petroselinum crispum* cell cultures. *Plant Cell Reports* 4, 277-280 (1985)
- Kombrink, E. and Hahlbrock, K. Rapid, systemic repression of the synthesis of ribulose 1,5-bisphosphate carboxylase small-subunit mRNA in fungus-infected or elicitor-treated potato leaves. *Planta* 181, 216-219 (1990)
- Latunde-Dada, A.O., Dixon, R.A. and Lucas, J.A. Induction of phytoalexin biosynthetic enzymes in resistant and susceptible lucerne callus lines infected with *Verticillium albo-atrum*. *Physiological and Molecular Plant Pathology* 31, 15-23 (1987)
- Li, J., Ou-Lee, T.M., Raba, R., Amundson, R.G. and Last, R.L. Arabidopsis flavonoid mutants are hypersensitive to UV-B irradiation. *Plant Cell* 5, 171-179 (1993)
- Ljung, L.J. and Ljung, E.J. *System identification: theory for the user*. New Jersey: Prentice-Hall (1998)
- Loake, G., Choudhary, A.D., Harrison, M.J., Mavandad, M., Lamb, C.J. and Dixon, R.A. Phenylpropanoid pathway intermediates regulate transient expression of a chalcone synthase gene promoter in electroporated protoplasts. *Plant Cell* 3, 829-840 (1991)
- Lois, R. Accumulation of UV-absorbing flavonoids induced by UV-B radiation in *Arabidopsis thaliana*. *Planta* 194, 498-503 (1994)
- Mandujano Chavez, A., Schoenbeck, M.A., Ralston, L.F., Lozoya Gloria, E. and Chappell, J. Differential induction of sesquiterpene metabolism in tobacco cell suspension cultures by methyl jasmonate and fungal elicitor. *Archives of Biochemistry and Biophysics* 381, 285-294 (2000)
- Matsumura, H., Nirasawa, S. and Terauchi, R. Transcript profiling in rice (*Oryza sativa* L.) seedlings using serial analysis of gene expression (SAGE). *Plant Journal* 20, 719-726 (1999)
- Maxwell, C.A., Harrison, M.J. and Dixon, R.A. Molecular characterization and expression of alfalfa isoliquiritigenin 2'-O-methyltransferase, an enzyme specifically involved in the biosynthesis of an inducer of *Rhizobium meliloti* nodulation genes. *Plant Journal* 4, 971-981 (1993)
- Mendes, P. GEPASI: a software package for modelling the dynamics, steady states and control of biochemical and other systems. *Computer Applications in the Biosciences* 9, 563-571 (1993)
- Mendes, P. Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends in Biochemical Sciences* 22, 361-363 (1997)
- Mendes, P. Metabolic simulation as an aid in understanding gene expression data. In *Workshop on Computation of Biochemical Pathways and Genetic Networks* (Bornberg-Bauer, E., De Beuckelaer, A. Kummer, U., Rost, U. eds) Berlin, Logos Verlag, pp. 27-34 (1999)

- Mendes, P. Modeling large biological systems from functional genomic data: parameter estimation. In *Foundations of Systems Biology* (Kitano, H. ed.) Cambridge MA, MIT Press, in press (2001)
- Mendes, P. and Kell, D.B. On the analysis of the inverse problem of metabolic pathways using artificial neural networks. *BioSystems* 38, 15-28 (1996)
- Mendes, P. and Kell, D.B. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics* 14, 869-883 (1998)
- Mitchell, D.L., Vaughan, J.E. and Nairn, R.S. Inhibition of transient gene expression in Chinese hamster ovary cells by cyclobutane dimers and (6-4) photoproducts in transfected ultraviolet-irradiated plasmid DNA. *Plasmid* 21, 21-30 (1989)
- Mueller, M.J., Brodschelm, W., Spannagl, E. and Zenk, M.H. Signaling in the elicitation process is mediated through the octadecanoid pathway leading to jasmonic acid. *Proceedings of the National Academy of Sciences USA* 90, 7490-7494 (1993)
- Mueller-Urli, F., Parthier, B. and Nover, L. Jasmonate-induced alteration of gene expression in barley leaf segments analyzed by in-vivo and in-vitro protein synthesis. *Planta* 176, 241-247 (1988)
- Muhlenbeck, U., Kortenbusch, A. and Barz, W. Formation of hydroxycinnamoyl amides and alpha-hydroxyacetovanillone in cell cultures of *Solanum khasianum*. *Phytochemistry* 42, 1573-1579 (1996)
- Ni, W., Fahrendorf, T., Ballance, G.M., Lamb, C.J. and Dixon, R.A. Stress responses in alfalfa (*Medicago sativa* L.). XX. Transcriptional activation of phenylpropanoid pathway genes in elicitor-treated cell suspension cultures. *Plant Molecular Biology* 30, 427-438 (1996a)
- Ni, W., Sewalt, V.J.H., Korth, K.L., Blount, J.W., Ballance, G.M. and Dixon, R.A. Stress responses in alfalfa (*Medicago sativa* L.) XXI. Activation of caffeic acid 3-O-methyltransferase and caffeoyl CoA 3-O-methyltransferase genes does not contribute to changes in metabolite accumulation in elicitor-treated cell suspension cultures. *Plant Physiology* 112, 717-726 (1996b)
- Nowacka, J. and Oleszek, W. High performance liquid chromatography of zanic acid glycosides in alfalfa (*Medicago sativa*). *Phytochemical Analysis* 3, 227-230 (1992)
- O'Farrell, P.H. High resolution two-dimensional electrophoresis. *Journal of Biological Chemistry* 250, 4007-4021 (1975)
- Oleszek, W. and Jurzysta, M. High-performance liquid chromatography of alfalfa root saponins. *Journal of Chromatography* 519, 109-116 (1990)
- Oleszek, W. Alfalfa saponins: structure, biological activity, and chemotaxonomy. In *Saponins Used in Food and Agriculture* (Waller and Yamasaki, eds), pp. 155-170, Plenum Press, New York (1996)
- Oleszek, W., Price, K. R. and Fenwich, G. R. Triterpene saponins from the roots of *Medicago lupulina* L. (black medic trefoil). *Journal of the Science of Food and Agriculture* 43, 289-291 (1988)
- Orr, J.D., Edwards, R. and Dixon, R.A. Stress responses in alfalfa (*Medicago sativa* L.) XIV. Changes in the levels of phenylpropanoid pathway intermediates in relation to regulation of L-phenylalanine ammonia-lyase in elicitor treated cell suspension cultures. *Plant Physiology* 101, 847-856 (1993a)
- Orr, J.D., Sumner, L.W., Edwards, R. and Dixon, R.A. Determination of cinnamic acid and 4-coumaric acid in alfalfa (*Medicago sativa* L.) cell suspension cultures by gas chromatography. *Phytochemical Analysis* 4, 124-130 (1993b)
- Paiva, N.L., Oommen, A., Harrison, M.J. and Dixon, R.A. Regulation of isoflavonoid metabolism in alfalfa. *Plant Cell, Tissue and Organ Culture* 38, 213-220 (1994)
- Paiva, N.L., Sun, Y., Dixon, R.A., VanEtten, H.D. and Hrazdina, G. Stress responses in alfalfa (*Medicago sativa* L.) XI. Molecular cloning and expression of alfalfa isoflavone reductase, a key enzyme of isoflavonoid phytoalexin biosynthesis. *Plant Molecular Biology* 17, 653-667 (1991)
- Pang, Q., Hays, J.B., Rajagopal, I. and Schaefer, T.S. Selection of Arabidopsis cDNAs that partially correct phenotypes of Escherichia coli DNA-damage-sensitive mutants and analysis of two plant cDNAs that appear to express UV-specific dark repair activities. *Plant Molecular Biology* 22, 411-426 (1993)
- Peltier, J-B., Friso, G., Kalume, D.E., Roepstorff, P., Nilsson, F., Adamska, I. and van Wijk, K.J., Proteomics of chloroplast: systematic identification and targeting analysis of lumenal and periferal thylakoid proteins. *Plant Cell* 12, 319-341 (2000).
- Pérez, N., Peña, S., Vega, S., Noa, M. and Enríquez, R. Medicagenic acid content in foliage of ten varieties of alfalfa (*Medicago sativa* L) cultivated in Mexico. *Journal of the Science of Food and Agriculture* 75, 401-404 (1997)
- Protic-Sabljić, M. and Kraemer, K.H. One pyrimidine dimer inactivates expression of a transfected gene in *Zeroderma pigmentosum*. *Proceedings of the National Academy of Sciences USA* 82, 6622-6626 (1986)
- Quinlan, J.R. *C4.5 Programs for machine learning*. San Mateo: Morgan Kaufmann (1993)

- Raventós, D., Jensen, A.B., Rask, M.B., Casacuberta, J.M., Mindy, J. and Segundo, B.S. A 20 bp *cis*-acting element is both necessary and sufficient to mediate elicitor response of a maize *PRms* gene. *Plant Journal* 7, 147-155 (1995)
- Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467-470 (1995)
- Schmelzer, E., Börner, H., Grisebach, H., Ebel, J. and Hahlbrock, K. Phytoalexin synthesis in soybean (*Glycine max*). Similar time courses of mRNA induction in hypocotyls infected with a fungal pathogen and in cell cultures treated with fungal elicitor. *FEBS Letters* 172, 59-63 (1984)
- Schumacher, H.-M., Gundlach, H., Fiedler, F. and Zenk, M.H. Elicitation of benzophenanthridine alkaloid synthesis in *Escholtzia* cell cultures. *Plant Cell Reports* 6, 410-413 (1987)
- Shorrosh, B.S., Dixon, R.A. and Ohlrogge, J.B. Molecular cloning, characterization and elicitation of acetyl CoA carboxylase from alfalfa. *Proceedings of the National Academy of Sciences USA* 91, 4323-4327 (1994)
- Siepel, A., Farmer, A., Tolopko, A., Zhuang, M., Mendes, P., Beavis, W. and Sobral, B. ISYS: A decentralized, component-based approach to the integration of heterogeneous bioinformatics resources. *Bioinformatics* in press (2000)
- Small, E. Adaptations to herbivory in alfalfa (*Medicago sativa*). *Canadian Journal of Botany* 74, 807-822 (1996)
- Sokal, R.R. and Michener, C.D. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* 28, 1409-1438 (1958)
- Somssich, I.E., Schmelzer, E., Bollmann, J. and Hahlbrock, K. Rapid activation by fungal elicitor of genes encoding "pathogenesis-related" proteins in cultured parsley cells. *Proceedings of the National Academy of Sciences USA* 83, 2427-2430 (1986)
- Steele, C.L., Gijzen, M., Qutob, D. and Dixon, R.A. Molecular characterization of the enzyme catalyzing the aryl migration reaction of isoflavonoid biosynthesis in soybean. *Archives of Biochemistry and Biophysics* 367, 147-150 (1999)
- Sumner, L.W., Paiva, N.L., Dixon, R.A. and Geno, P.W. High-performance liquid chromatography/continuous-flow liquid secondary ion mass spectrometry of flavonoid glucosides in leguminous plant extracts. *Journal of Mass Spectrometry* 31, 472-485 (1996)
- Swayne, D.F., Cook, D. and Buja, A. XGobi: interactive dynamic data visualization in the X window system. *Journal of Computational and Graphical Statistics* 7, 113-130 (1998)
- Takayanagi, S., Trunk, J.G., Sutherland, J.C. and Sutherland, B.M. Alfalfa seedlings grown outdoors are more resistant to UV-induced damage than plants grown in a UV-free environmental chamber. *Photochemistry and Photobiology* 60, 363-367 (1994)
- Tava, A. and Odoardi, M. Saponins from *Medicago* spp: chemical characterization and biological activity against insects. In *Saponins Used in Food and Agriculture* (Waller and Yamasaki, eds), Plenum Press, New York (1996)
- Thiellement, H., Bahrman, N., Damerval, C., Plomion, C., Rossignol, M., Santoni, V., de Vienne, D. and Zivy, M. Proteomics for genetic and physiological studies in plants. *Electrophoresis* 20, 2013-2026 (1999)
- Trethewey, R.N., Krotzky, A.J. and Willmitzer, L. Metabolic profiling: a Rosetta Stone for genomics? *Current Opinion in Plant Biology* 2, 83-85 (1999)
- Trieu, A.T., Burleigh, S.H., Kardailsky, I.V., Maldonado-Mendoza, I.E., Versaw, W.K., Blaylock, L.A., Shin, H., Chiou, T.-J., Katagi, H., Dewbre, G.R., Weigel, D. and Harrison, M.J. Transformation of *Medicago truncatula* via infiltration of seedlings or flowering plants with *Agrobacterium*. *Plant Journal*, in revision (2000)
- van de Löcht, U., Meier, I., Hahlbrock, K. and Somssich, I.E. A 125 bp promoter fragment is sufficient for strong elicitor-mediated gene activation in parsley. *EMBO Journal* 9, 2945-2950 (1990)
- Vonarx, E.J., Mitchell, H.A., Karthikeyan, R., Chatterjee, I. and Kunz, B. DNA repair in higher plants. *Mutation Research* 400, 187-200 (1998)
- Wagoner, W., Loschke, D.C. and Hadwiger, L.A. Two-dimensional electrophoretic analysis of *in vivo* and *in vitro* synthesis of proteins in peas inoculated with compatible and incompatible *Fusarium solani*. *Physiological Plant Pathology* 20, 99-107 (1982)
- Wahde, M. and Hertz, J. Coarse-grained reverse engineering of genetic regulatory networks. *BioSystems* 55, 129-136 (2000)
- Watson, D.G. and Pitt, A.R. Analysis of flavonoids in tablets and urine by gas chromatography/mass spectrometry and liquid chromatography/mass spectrometry. *Rapid Communications in Mass Spectrometry* 12, 153-156 (1998)

- Wolf, B.P., Sumner, L.W., Campbell, J.M., DeGnore, J.P. and Juhasz, P., Utilization of MALDI-TOF-TOF technology for the homology comparison and sequencing of *Medicago truncatula* plant proteins. *Proceedings of the 46th ASMS Conference on Mass Spectrometry and Allied Topics*, Palm Springs, CA (2000).
- Wolf, B.P., Sumner, L.W., Shields, S.J., Nielsen, K., Gray, K.A. and Russell, D.H. Characterization of proteins utilized in the desulfurization of petroleum products by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *Analytical Biochemistry* 260, 117-27 (1998)
- Yang, P.Z., Chen, C.H., Wang, Z.P., Fan, B.F. and Chen, Z.X. A pathogen- and salicylic acid-induced WRKY DNA-binding activity recognizes the elicitor response element of the tobacco class I chitinase gene promoter. *Plant Journal* 18, 141-149 (1999)
- Yates, J.R. III. Mass spectrometry and the age of the proteome. *Journal of Mass Spectrometry* 33, 1-19 (1998)
- Yu, L.M., Lamb, C.J. and Dixon, R.A. Purification and biochemical characterization of two proteins which bind to the H-box *cis*-element implicated in transcriptional activation of plant defense genes. *Plant Journal* 3, 805-816 (1993)
- Zenk, M.H. Chasing the enzymes of secondary metabolism: plant cell cultures as a pot of gold. *Phytochemistry* 30, 3861-3863 (1991)