Plant Physiology Preview. Published on November 13, 2009, as DOI:10.1104/pp.109.149583

Bioinformatic and systems biology tools to generate testable models of signalling pathways and their targets

Andrea Pitzschke[a] and Heribert Hirt[b,c]

[a]Department of Applied Genetics and Cell Biology; University of Natural Resources and Applied Life Sciences, Muthgasse 18, 1190 Vienna, Austria
[b]Department of Plant Molecular Biology, Max F. Perutz Laboratories, University of Vienna, Dr.-Bohr-Gasse 9, 1030 Vienna, Austria
[c]URGV Plant Genomics Laboratory, 2 Rue Gaston Cremieux, 91057 Evry, France

Abstract

Over the last years a number of bioinformatic software programs have been developed in the area of molecular biology. The application of these bioinformatic tools to the wealth of existing transcriptomic and proteomic data can be used to predict the structure and hierarchy of signalling pathways and gene networks. In genetically tractable model organisms such as *Arabidopsis thaliana*, these hyphotheses can be validated experimentally and modified in reiterative cycles, giving hypothesis-driven research high feasibility. These predictive systems biology approaches significantly reduce the scale, time and manpower usually required in classical approaches. Here, we provide an overview on the use of currently available tools in deciphering signalling pathways in Arabidopsis research.

Introduction

With more and more high-throughput data becoming available, scientists are faced with the challenge to develop or apply intelligent software to extract essential information from large-scale data sets. If used in a “smart way”, some bioinformatic programs can aid in many ways to elucidate the function of a gene of interest, including modes of regulation and synthesis, its post-translational modifications and potential interaction partners and last not least processes that are regulated by its gene products. Examples of combinatory applications of bioinformatic tools that lead to the generation (and subsequent confirmation) of hypotheses are described below; with a focus on the deciphering of cellular processes regulated by MAPK cascades in Arabidopsis.

Plants need to cope with a wide range of challenging environmental conditions. The successful adaptation/response to such stresses requires the efficient and specific transduction of environmental signals. In stress signal transduction, a prominent role is played by MAPK

cascades, which minimally consist of a MAPKKK, a MAPKK and a MAPK. Via a phosphorelay mechanism, these modules transduce incoming signals to activate MAPKs which subsequently phosphorylate specific target proteins (reviewed in Colcombet and Hirt, 2008; Pitzschke et al., 2009a). So far, experimental evidence exists only for a very few MAPK substrates, but a proteomic phosphoarray approach suggests that transcription factors (TFs) are the major targets of MAPKs (Popescu et al., 2009). Phosphorylation of TFs can potentially alter their subcellular localisation, protein stability or DNA-binding activity. MAPK cascades may thus be primary regulators of stimulus-dependent adaptation of gene expression. The Arabidopsis genome encodes for 60-80 MAPKKKs, 10 MAPKKs and 20 MAPKs. The present challenge is to elucidate i) which of the thousands of theoretically possible signalling modules are indeed formed, ii) which stimuli are conveyed, iii) which targets are addressed, and iv) what is the biological role of the respective signalling modules.

Comparison of expression profiles

One approach is to use correlative transcriptome analysis as a relatively unbiased technique. Hereby, microarray profiles of signalling cascade mutants can be compared from a wide range of organisms, e.g. by using the Genevestigator tool https://www.genevestigator.com/gv/index.jsp. Mutants whose transcriptome profiles significantly overlap are likely to act in common signalling cascades. The extent of such overlaps can be conveniently visualised in Venn diagrams, e.g. using the tool at http://www.pangloss.com/seidel/Protocols/venn4.cgi, where expression profiles of up to four mutants can be compared. The program also generates lists of gene IDs occurring in 2, 3 or 4 entered data sets. Further inspection of the list of commonly regulated genes can give indications on the processes controlled by a theoretical signalling module, e.g. using Genevestigator - a rich source for transcriptome data on spatio-temporal expression patterns, mutant profiles and responses to numerous treatments/growth conditions.

The following example emphasises the consistencies with respect to similarities in expression profiles, phenotype and hormone accumulation and thus documents the robustness and usefulness of transcriptome-based approaches. Rather than confirming correlations predicted from experimental results, with comparatively little effort such tools can generate reasonable hypotheses, which can subsequently be experimentally validated.

Example: MEKK1-MKK1/2-MPK4 and beyond

MEKK1-MKK1/2-MPK4 engage in a signalling cascade that is activated in response to pathogen attack (Gao et al., 2008). Bimolecular fluorescence complementation (BiFC) analysis showed that both MPK4 and MEKK1 interact with MKK1 and MKK2 (Gao et al., 2008). *mekk1*, *mpk4* and *mkk1/mkk2* double knock out mutants show spontaneous cell lesions and highly elevated levels of reactive oxygen species. Moreover, they display a severely dwarfed phenotype, which is correlated with the strong accumulation of salicylic acid (SA), a major hormone in biotic pathogen responses. Accordingly, the sensitivity to the plant pathogen *Pseudomonas syringae* is reduced in these pathway mutants. For these reasons, the MEKK1-MKK1/2-MPK4 cascade has been ascribed a role as a negative regulator of innate immune responses in plants (Gao et al., 2008).

For all mutants affected in this MAPK module, transcriptome analyses have been performed (Qiu et al., 2008a; Pitzschke et al., 2009b). Indeed, the gene expression profiles of these mutants are highly similar. Consistent with the hierarchical order in the signalling cascade, *mekk1* show the largest set of differentially regulated genes, followed by *mkk1/2* and eventually *mpk4*. Moreover, many of the common differentially regulated genes are known to be SA-responsive genes and/or are associated with redox regulation (Qiu et al., 2008a; Pitzschke et al., 2009b). In agreement with the partial redundancy of MKK1 and MKK2, the expression profiles of *mkk1* or *mkk2* single mutants hardly overlap with those of *mekk1*, *mkk1/mkk2* and *mpk4* mutants (Gao et al,. 2008). Microarray data-based online tools (e.g. https://www.genevestigator.com/gv/index.jsp) also reveal a strong correlation of transcriptome profiles of *mekk1, mkk1/mkk2* and *mpk4* with several other mutants, such as *cpr5* (*constitutive expression of PR genes 5*) and *npr1* (*non-expressor of pathogenesis-related genes 1*), suggesting further commonalities between these mutants. A more targeted bioinformatic approach of comparative transcriptome studies, Functional Associations of Response Overlap (FARO), has also highlighted the relatedness of *mekk1, mkk1/2, mpk4* profiles with those of *cpr5* and *npr1* (Nielsen et al., 2007). Indeed, *cpr5* and *npr1* mutants are also dwarfed and have highly elevated SA levels (Bowling et al., 1997; Cao et al., 1994).

An interesting characteristic of many SA-accumulating mutants is that their dwarfed phenotype (often associated with sterility/poor seed production) can be rescued by growing these plants at elevated temperatures, in line with the observed negative correlation of the heat-induced and the *mpk4* mutant gene expression profiles (revealed by FARO analysis, Nielsen et al., 2007). Applying this knowledge to lines of interest may assist the positioning of the corresponding gene in the signalling cascade and can help to yield a larger pool of precious seeds through adjustment of growth conditions.

Web-based tools that integrate large sets of microarrays have the potential to reveal novel correlations between responses. To give an example, we observe a strong negative correlation between the expression response to SA and CO_2 by Genevestigator analysis. It may therefore be worth testing the CO_2 response of *mekk1*, *mkk1/2* and *mpk4* mutants with respect to phenotype, SA-levels and transcriptional changes. Likewise, this observation may also indicate that increasing environmental pollution (CO_2) renders plants more susceptible to pathogen attack and a recent study provides experimental evidence for this in silico-based assumption (Lake and Wade, 2009).

How to GO from gene to function

To understand the functional significance of gene expression profiles displayed by a mutant of interest, a search for statistically overrepresented "functional " and "cellular compartment" terms, using the gene ontology (GO) tool (e.g. http://www.arabidopsis.org/tools/bulk/go/index.jsp) is another promising approach. Not unexpectedly, in our example, this tool detects an enrichment of the GO terms "stress-responsive" and "transcription factor activity" in the list of *mekk1*, *mkk1/2* and *mpk4*-upregulated genes. Moreover, GO term analysis revealed that among the genes down-regulated in *mekk1* and *mpk4* those encoding plastidic or chloroplastic proteins are significantly overrepresented (Pitzschke et al., 2009b), which may indicate that these mutants might also regulate processes related to photosynthesis in order to prevent further ROS production.

How to identify potentially co-regulated genes - the Arabidopsis Chromosome Map tool

The highly user-friendly setup and the diversity of tools provided by TAIR enables the researcher to subject genes of interest to further bioinformatic analysis. For example, the tool (http://www.arabidopsis.org/jsp/ChromosomeMap/tool.jsp), which displays the position of entered genes on the 5 Arabidopsis chromosomes, is useful for revealing clustering of genes on a particular chromosomal region. Such local clustering can be an indication for transcriptional co-regulation, as e.g. evidenced in a study on the cluster of Arabidopsis *RPP5* locus *R* genes involved in pathogen response (Yi and Richards, 2007).

How to manage the flood of data?

Despite their unquestionable value for driving research progress, whole-genome microarrays have their drawbacks. The experiment as such is very cost-intensive. Moreover, the huge

dataset generated from these arrays often confronts the scientist with the (hard) decision which subset of differentially expressed genes to investigate further (*in silico*). Prior selection might therefore be advisable. For those researchers particularly interested in defence-related responses, the small-scale expression array (Sato et al, 2007), which analyses transcript abundance of 321 genes associated with pathogen response (and a set of genes for normalisation), may be a good alternative, either for the experiment as such and/or as a pre-selection for the downstream data analysis. Results from a recent miniarray study investigating the response of seven defence-affected mutants (*coi1*, *dde2*, *ein2*, *mpk3*, *pad4*, *cbp60g-1*, *sid2*) to *P. syringae* treatment (Wang et al., 2009) provides a manageable data set for a first comparison with own data. The closer the transcriptome profile of a mutant of interest is to any of these mutants, the higher the probability that the corresponding proteins engage in a common pathway. Moreover, if a mutant profile shows strongest overlap with the subset of genes co-regulated in several of the other mutants (e.g. subsets of the seven defence-related mutants), the corresponding protein may be an upstream regulator acting before stress signalling bifurcation into individual pathways.

Similar to the usefulness of the above-described miniarray for pathogen response-focussed research, a global map of gene expression within 15 different zones of the root corresponding to cell types and tissues at progressive developmental stages, allows researchers a pre-selection of large datasets (retrieved from published or own microarrays) for the analysis of developmental aspects (Birnbaum et al., 2004). Likewise, a report from Leonhardt et al. (2004) provides a list of genes that allows a pre-selection for guard cell- and mesophyll-expressed genes.

Mapping of microarray data onto pathways and genetic maps – The MapMan tool

One precious tool, which allows analysis of large datasets and facilitates the assignment of clusters of genes showing major transcriptional changes to areas of function, is MapMan (http://gabi.rzpd.de/projects/MapMan). MapMan is grouping genes on the Arabidopsis affymetrix 22K array into >200 hierarchical categories, thereby providing an overview of various cellular processes. Due to its complexity, we will not describe this tool in detail, but recommend the following articles (Usadel et al., 2005; Thimm et al., 2004). Ideally, upon reading of the articles, MapMan should be visited and data sets, including those of own experiments, explored.

Briefly, MapMan allows superimposition of different datasets in overlay plots and thus facilitates the identification of shared features, both globally and on a gene-to-gene basis.

By grouping genes that are probably involved in a common area of function, the MapMan tool can reveal trends towards repression or induction, which might not be obvious at the single gene level. The datasets of responses of interest can originate from own experiments or can be downloaded from published microarrays. The analysis is also facilitated by the option to focus and visualise certain major pathways, such as “metabolism” or “DNA synthesis”.

The usefulness of MapMan has been demonstrated by the analysis of the Arabidopsis starvation response: The transcript profile of wild type seedlings harvested at the end of the night was compared either to wild type seedlings that had been incubated in the dark for an additional 6-hour-period or to starchless *pgm* mutants harvested at the end of the night. The MapMan-generated overlay plot revealed strong correlation between these two sugar-depletion conditions. As might be expected, the common transcriptional response indicates repression of photosynthesis and sucrose, starch and lipid synthesis, while genes involved in lipid, amino acid and carbohydrate breakdown are largely induced. Novel aspects of sugar depletion were also revealed, e.g. a trend to preferential induction of cell-wall synthesis-involved genes and repression of genes involved in cell wall breakdown. Furthermore, previous indications on a cross-talk between sugar sensing and ABA- and ethylene sensing pathways (Brocard et al., 2002; Leon and Sheen, 2003; Rook et al., 2001) could be substantiated. MapMan is being up-dated continuously, and a conversion of this tool now also allows comparison of responses in different organisms, as demonstrated by the comparison of diurnal changes in Arabidopsis and tomato expression profiles (Urbanczyk-Wochniak et al., 2006).

Despite its unquestionable value, MapMan has the major drawback in that many genes cannot be categorized into certain MapMan-defined areas of function and are therefore not considered in the analysis. For example, in a study on the Arabidopsis response to *Fusarium*, the majority of genes could not be assigned to any of the known function categories (Yuan et al., 2008). If one´s list of genes of interest contains several “genes of unknown function” a further separate inspection might be advisable. Using ClustalW (http://www.genebee.msu.su/clustal/basic.html) potential phylogenetic relatedness between the corresponding proteins can be detected, which will help to assign putative roles/implications of those proteins to the process that is investigated. This, in turn, can help to refine the MapMan datasets and thus facilitate future analyses.

Prediction of pathway modules through correlation of gene expression

Genes that are co-expressed over multiple data sets are likely to show functional relatedness. This knowledge may help to predict which proteins act in a common pathway or – as in this particular case – which MAPK signalling component engages in a common module: Using the *AttedII tool* (http://atted.jp/), lists of genes whose expression correlates with that of a gene of interest can be generated and correlation coefficients calculated. To test its suitability, we queried *AttedII* to predict components potentially associated with MKK4, a stress-related MAPKK whose transcript abundance alters in response to numerous stimuli (as e.g. evidenced in Genevestigator). *AttedII* reveals strong gene expression correlation of *MKK4* with *MKK5* and also with *MPK3*, but not with any other MAPK signalling component. MKK4 and MKK5 are known to be functionally redundant, to be controlled by the MAPKKK YODA, and to act as upstream regulators of the MAPKs MPK3 and MPK6 (Wang et al., 2008). Neither *YODA* nor *MPK6* are among the predicted *MKK4*-correlated genes, most likely due to their ubiquitous expression. Further genes correlating with *MKK4* expression are promising candidates for encoding additional components involved in MKK4-mediated signal transduction. The above example shows the usefulness of gene expression correlation-based hypothesis generation, but also reveals the limitations of this approach for constitutively expressed genes.

How to exploit microarray data for the identification of targets of signalling cascades?

The rich pool of publicly available microarray data cannot only be screened by bioinformatic tools for hypothesising the composition of signalling pathways, but they are also suitable for making predictions on the transcription factors and promoter elements that control a set of co-expressed genes.

Although each type of signal requires a specific cellular response, the transcript abundance of some genes is altered in response to multiple signals. This approach can be exemplified for finding the set of common stress genes by using a clustering method (Ma and Bohnert, 2007). Using publicly available microarray data of transcriptional changes in response to various abiotic and biotic stresses, 197 common stress responsive genes were identified. Similar studies were reported by Swindell et al. (2006), and by Kant et al. (2008) for 9 and 16 abiotic stress conditions, respectively. Based on GO annotation (kinase, TF, etc.), the latter report classified a subset of 289 genes as multiple stress regulatory genes (MSTRs), including several members of the WRKY and bZIP protein families, which are known to be stress associated (reviewed in Ülker and Somssich, 2004; Jakoby et al., 2002). Considering the transcriptional response of these factors to very diverse signals, one may position MSTRs at

the early steps of stress signalling responses. MSTRs can be expected to have a high turnover and to be controlled at multiple levels in order to allow fast adaptation and to prevent a prolonged activation of downstream signalling processes that would interfere with plant growth and development. These characteristics render MSTRs prime candidates for post-translational modifications such as protein kinases or ubiquitin-mediated stability control.

Given that some signalling cascades are activated very rapidly (e.g. MAPK cascades are activated within minutes), candidate targets for diverse signalling pathways might be found by screening transcriptome data sets for very early responses in a similar fashion as done for MSTRs.

The identification of early targets of signalling cascades and knowledge on the modes controlling their activity is also of tremendous value for applied science, because appropriate manipulation may minimize the effort of creating crops with desired traits such as resistance to multiple stresses. The key regulators may be expressed in a controllable system, e.g. by using chemically inducible expression or nuclear translocation systems, thereby circumventing undesirable side effects on growth/development that are often associated when over-expressing genes constitutively.

How to exploit microarray data for the identification of transcriptional regulators?

In order to decipher the modes controlling the expression of a set of co-expressed genes, an in-depth inspection of their upstream regulatory regions may provide further information. During the immediate responses to a given stimulus, the signalling may not have bifurcated yet into highly complex downstream pathways. Therefore, early-induced genes are likely to underlie regulation by common TF(s) and therefore share common DNA motifs in their regulatory regions. Promoter sequences of user-defined length can be downloaded from several Arabidopsis databases, e.g. TAIR (http://www.arabidopsis.org/tools/bulk/sequences/index.jsp), and subsequently screened for the presence of certain DNA motifs. While PLACE (www.dna.affrc.go.jp/PLACE/)or PlantCARE (http://sphinx.rug.ac.be:8080/PlantCARE/) are useful for detecting known cis-elements within a set of promoters, the TAIR motif finder (http://www.arabidopsis.org/tools/bulk/motiffinder/index.jsp) and AlignACE tool (http://atlas.med.harvard.edu/cgi-bin/alignace.pl) allow identification of potentially novel DNA motifs shared by multiple promoters. Once candidate motifs have been identified, the statistical significance of their enrichment can be assessed using the POBO tool (http://ekhidna.biocenter.helsinki.fi/poxo/pobo/pobo), which compares motif abundance in the

given promoter set to the Arabidopsis background frequencies. This tool has e.g. proven useful for documenting the strong enrichment of W-boxes in the promoters of WRKY18-dependent SA-inducible genes (Wang et al., 2006).

Subsequent to this statistical analysis, the functional relevance of enriched candidate DNA motifs in mediating stress responses can be experimentally validated using synthetic promoter-reporter gene constructs in transgenic plants or transfected protoplasts. The latter system also allows - with minimal effort - to test candidate transcription factors for their ability to induce/repress gene expression driven by a motif of interest, as e.g. evidenced in Rushton et al. (2002) or Pitzschke et al. (2009c).

How to find the targets of transcription factors?

Alternatively to starting with the identification of multiple signal-responsive genes through the comparison of multiple signal-dependent expression profiles, an equally attractive approach for the elucidation of signalling cascades is the detailed characterisation of a TF of interest, e.g. a known or predicted substrate of a signalling cascade.

For their characterisation, a phylogenetic analysis may provide first indications about the dimerisation behaviour and sometimes even about potential DNA target motifs. However, high homology within the DNA binding domain of two TFs does not necessarily correlate with target motif similarity. For example, the bZIP factors tobacco RSG2 and tomato VSF-1 have a highly conserved bZIP domains, yet they bind to completely different DNA motifs. (Fukazawa et al., 2002; Ringli et al., 1998). The bZIP domain of Arabidopsis VIP1 is strongly related to those of RSG2 and VSF-1 and VIP1 had been shown earlier to be phosphorylated by MPK3 in a stress-dependent manner and to undergo cytoplamic-nuclear translocation (Djamei et al., 2007). Where no further information on the DNA motifs targeted by a TF of interest is available, random DNA selection assays (RDSA) may be applied to generate data that subsequently can be analysed by a range of bioinformatic tools (Pitzschke et al., 2009c).

In RDSA, random double-stranded DNA fragments, usually 15-20 nucelotide long and flanked by defined primer-annealing sites, are incubated with recombinant TF protein. Candidate motifs are enriched through a repetitive selection – amplification procedure. RDSA yields a range of candidate DNA motifs which can be screened for common elements and aligned using the STAMP tool (http://www.benoslab.pitt.edu/stamp/). Electrophoretic mobility shift assays and mutagenesis of the candidate motif(s) is then used for confirming the binding and specificity of the TF to those motifs. Once such motif has been found and confirmed, target genes of the TF can be predicted. For this, the TAIR patmatch tool

(http://www.arabidopsis.org/cgi-bin/patmatch/nph-patmatch.pl) provides a tab-delimited file of position, number and orientation for all genes harbouring such motifs in a user-defined region (e.g. within 500 bp promoter regions). In the case of VIP1, this information aided the prediction of one of its target genes *MYB44*, which was later confirmed by promoter-reporter gene activation and chromatin immunoprecipitation (Pitzschke et al., 2009c). For several multimeric TFs the spacing between adjacent target DNA motifs is crucial for transactivating activity. If knowledge about the spacing exists (e.g. the preferred spacing between W-boxes targeted by certain groups of WRKYs (Ciolkowski et al., 2008), the number of further candidate target genes can be narrowed down. A fast visual tool for this application is the MotifMatcher tool (http://users.soe.ucsc.edu/~kent/improbizer/motifMatcher.html), which depicts multiple user-defined motifs (entered as matrix), each in a different colour, on a set of promoters of interest as "beads on a string" (Figure 1).

How to use proteomic data in constructing signalling networks?

Hypotheses on signalling pathway compositions cannot only be generated through gene expression-based analyses, but also through proteomic approaches. The "classical" experimental approach for retrieving a list of candidate interactors of a protein of interest are yeast-two-hybrid (Y2H) screens and mass spectrometry (MS) analysis of purified protein complexes. Whereas Y2H analyses have the potential to predict direct protein interaction partners, MS analysis of protein complexes primarily indicates that the proteins are in more or less complicated assemblies of proteins. The low degree of overlap in Y2H and MS studies in yeast further cautions on a naïve interpretation of these data sets. Y2H suffer from a relatively high degree of false positives that can be generated by multiple factors that are inherent in the system, including overexpression, artificial interaction of two components in the same compartment or misfolding of the protein of interest due to fusion to yeast bait or prey proteins, respectively. MS studies of protein complexes, on the other hand, suffer from co-purification of more or less abundant contaminants and the possibility that the proteins may not be interacting directly. Given these draw backs, valuable information can nonetheless be obtained from *in silico* analysis of publicly available interaction data, e.g. by using the tool provided at (http://bar.utoronto.ca/interactions/cgi-bin/arabidopsis_interactions_viewer.cgi). This tool queries a huge database of confirmed Arabidopsis interacting proteins retrieved from Biomolecular Interaction Network Database and from high-density Arabidopsis protein microarrays, and provides details about the experimental evidence. It also integrates data from

macro- and microarray-based phosphoprotein arrays that led to the identification of Arabidopsis MAPK candidate substrates (Feilner et al., 2005; Popescu et al., 2009).

Because a Y2H- or protein microarray-based predicted interaction does not necessarily mean that two proteins truly interact *in planta*, the list of candidate interacting proteins can be narrowed down by applying additional selection criteria:

a) check the spatio-temporal expression pattern of the corresponding genes ("Does gene x expression overlap with that of gene y"). Useful tools are: (https://www.genevestigator.com/gv/index.jsp or http://atted.jp/).

b) compare the subcellular localisation of the proteins (a chloroplast-localised protein is unlikely to interact with a nuclear protein).

Obviously, data merely based on prediction algorithms (e.g. http://wolfpsort.org//; http://www.cbs.dtu.dk/services/TargetP/) need to be interpreted with caution. The more complex SUBA tool (http://www.plantenergy.uwa.edu.au/suba2/) integrates prediction-based information and data based on experimental evidence (MS/MS, GFP fusion protein localisation studies). Through integration of transcriptomic and proteomic data – from own and published arrays - can also further facilitate the identification of "top candidates " (figure 2). Once a list of a manageable number of candidate interaction partners has been established, their ability to bind to the protein of interest can be experimentally validated (co-immunoprecipitation/BiFC/FRET).

The "interaction viewer" tool and the screening of published lists of protein-protein interactions can also aid the prediction of (further) partners that interact with a protein of interest. (If protein A interacts with B, and B with C, then A might also interact with C). For instance, MPK4 has been shown to interact with MKS1 (MAPK substrate 1). On the other hand, MKS1 was found to interact with two WRKY transcription factors, WRKY25 and WRKY33 in yeast. Both WRKYs are involved in biotic stress signalling, which in turn is clearly linked to MPK4. In an elegant series of experiments, Qiu et al. (2008b) could show that MPK4 exists in nuclear complexes with the WRKY33 transcription factor. This complex depends on the MPK4 substrate MKS1. Challenge with pathogenic elicitors leads to the activation of MPK4 and phosphorylation of MKS1. Subsequently, complexes with MKS1 and WRKY33 are released from MPK4, and WRKY33 is recruited to the promoter of PHYTOALEXIN DEFICIENT3 (PAD3) encoding an enzyme required for the synthesis of antimicrobial camalexin. MKS1 serves to fine-tune WRKY33-mediated PAD3 expression. In line with this scenario, *wrky33* mutants exhibit enhanced susceptibility to necrotrophic

pathogens, whereas overexpression of *WRKY33* increases resistance (Zheng et al, 2006, 2007). A recent transcriptome study has revealed further potential target genes of WRKY33, including *CYP71A1* which encodes a cytochrome P450 monoxygenases required for camalexin synthesis (Petersen et al., 2008). The AttedII tool predicts a strong co-regulation of *PAD3* with *CYP71A13*, and the promoters of both genes carry multiple W-boxes, suggesting that both genes underlie a common regulatory mechanism, i.e. through WRKY33 (Petersen et al., 2008).

The in-depth study of the precious pool of protein sequences may aid the prediction of peptide motifs that are recognised by a given kinase. Moreover, similar to the screening of Arabidopsis genes carrying a motif of interest in their upstream regulatory region, the TAIR patmatch tool can be applied to generate a list of candidate Arabidopsis proteins that harbour a given peptide motif. Additional confidence about the functional relevance of a candidate peptide motif may also be obtained through phylogenetic analysis. E.g. Arabidopsis NPR1 and its orthologs in other plants carry a characteristic DSXXXS peptide, "phosphodegron", which marks it for phosphorylation-dependent proteasomal degradation (Spoel et al., 2009). Phylogenetic analysis can e.g. be performed using the tool at http://bioinfoserver.rsbs.anu.edu.au/utils/affytrees/, which provides information about the homologs to a protein of interest in other plant species.

The functional relevance of candidate peptide motifs can then be experimentally verified (e.g. through *in vitro* phosphorylation). Subsequently, hybrid/artificial kinases can be created that modify proteins other than their true targets or that prevent phosphorylation of a protein by outcompeting the "true" modifying upstream kinase. Given that phosphorylation events are a common feature in the signalling of critical responses/processes in animals, this approach has high potential e.g. for tumorigenesis/cancer therapy research.

In summary, this review documents the usefulness, robustness and limitations of applying various transcriptome-, promoterome- and proteome-based bioinformatic tools for deciphering signalling pathways in Arabidopsis. Clearly, only a small subset of available tools are described, and their -literally- unlimited number of elaborate combinations harbours high potential to significantly speed up the progress in signalling research. Also, modelling approaches, e.g. based on kinetic data, harbour a huge potential to dissect signalling pathways. In the long run, in silico analysis will replace bench work to a large extent, because experiments can be designed in a highly targeted manner.

List of bioinformatic tools described in this review

tool	link	description	ref.
DNA tools			
Genevestigator	https://www.genevestigator.com/gv/index.jsp	study the expression and regulation of genes in a broad variety of contexts.	Zimmermann et al., 2004
AttedII	http://atted.jp/	find coexpressed genes	Obayashi et al., 2007
TAIR bulk sequence download	http://www.arabidopsis.org/tools/bulk/sequences/index.jsp	bulk download of sequences (transcript, regulatory regions, UTR, introns...) for lists of gene IDs of interest.	Garcia-Hernandez et al., 2002
TAIR patmatch	http://www.arabidopsis.org/cgi-bin/patmatch/nph-patmatch.pl	pattern matching tool to search for short (<20 residues) nucleotide or peptide sequences, or ambiguous/degenerate patterns in Arabidopsis DNA/protein sequences.	Yan et al., 2005
PLACE	www.dna.affrc.go.jp/PLACE/ http://sphinx.rug.ac.be:8080/PlantCARE/	detect known cis-elements within a promoter set of interest	Higo et al., 1998; Lescot et al., 2002
TAIR motif finder AlignACE	http://www.arabidopsis.org/tools/bulk/motiffinder/index.jsp http://atlas.med.harvard.edu/cgi-bin/alignace.pl	detect novel cis-elements within a promoter set of interest	
STAMP	http://www.benoslab.pitt.edu/stamp/	alignment, similarity, and database matching for DNA motifs	Mahony et al., 2007
POBO	http://ekhidna.biocenter.helsinki.fi/poxo/pobo/pobo	calculate and visually display the significance of putatively enriched DNA elements	Kankainen and Holm, 2004
MotifMatcher	http://users.soe.ucsc.edu/~kent/improbizer/motifMatcher.html	visualise user-defined multiple DNA motifs on promoter sets of interest as "beads on a string".	
MapMan	http://gabi.rzpd.de/projects/MapMan	Map transcript profiling data onto pathways and onto genetic maps; generate response overlays	Thimm et al., 2004; Usadel et al., 2005
protein tools			
TAIR patmatch	http://www.arabidopsis.org/cgi-bin/patmatch/nph-patmatch.pl	retrieve lists of gene IDs containing a (<20 residues) nucleotide or peptide sequence of interest	Garcia-Hernandez et al., 2002
BAR interactome	http://bar.utoronto.ca/interactions/cgi-bin/arabidopsis_interactions_viewer.cgi	find interacting proteins 70,000 predicted protein–protein interaction data by Geisler-Lee et al. (2007) and 2800 documented *Arabidopsis* protein–protein interactions.	Geisler-Lee et al., 2007
TargetP WolfPSort	http://www.cbs.dtu.dk/services/TargetP/ http://wolfpsort.org/	subcellular localisation (prediction algorithms)	Horton et al., 2007

SUBA	http://www.plantenergy.uwa.edu.au/suba2/	subcellular localisation (from prediction algorithms and experimental evidence)	Heazlewood et al., 2007
Affytree	http://bioinfoserver.rsbs.anu.edu.au/utils/affytrees/	phylogeny tool	Frickey et al., 2008
GenebeeClustalW	www.genebee.msu.su/clustal	alignment and phylogenetic analysis of amino acid sequences	Frickey et al., 2008
classification tools			
TAIR GO annotation	http://www.arabidopsis.org/tools/bulk/go/index.jsp	GO annotation search, functional categorization	Garcia-Hernandez et al., 2002
TAIR chromosome map tool	http://www.arabidopsis.org/jsp/ChromosomeMap/tool.jsp	Displays entered geneIDs on the Arabidopsis chromosomes	Garcia-Hernandez et al., 2002
Venn generator	http://www.pangloss.com/seidel/Protocols/venn4.cgi	Venn diagram generator. Up to four lists of geneIDs can be analysed.	

Figure legends

Figure1

Flow chart of a possible strategy to identify the DNA motif(s) and promoters targeted by a TF of interest. Bioinformatic tools are shown in italics.

Figure 2

Prediction of signalling components through integration of transcriptome and proteome arrays. By searching for the overlap between candidate proteins identified from peptide-based microarrays (e.g. interactors of a regulatory protein in a process of interest; right) and proteins encoded by genes differentially expressed under conditions of interest (left), high-priority candidates involved in the immediate downstream signalling can be defined. As exemplified here, multiple-stress-responsive genes that encode for proteins for which phosphorylation by a stress-associated kinase/phosphatase has been observed, are likely to be key components in early stress signal transduction.

References

Bowling SA, Clarke JD, Liu Y, Klessig DF, Dong X (1997) The cpr5 mutant of Arabidopsis expresses both NPR1-dependent and NPR1-independent resistance. Plant Cell **9:** 1 573-1584

Birnbaum K, Jung JW, Wang JY, Lambert GM, Hirst JA, Galbraith DW and Benfey PN (2005) Cell type–specific expression profiling in plants via cell sorting of protoplasts from fluorescent reporter lines. Nature Methods **2**: 615 - 619

Brocard IM, Lynch TJ, Finkelstein RR (2002) Regulation and role of the Arabidopsis abscisic acid-insensitive 5 gene in abscisic acid, sugar, and stress response. Plant Physiol **129:** 1533-1543

Cao H, Bowling SA, Gordon AS, Dong X (1994) Characterization of an Arabidopsis Mutant That Is Nonresponsive to Inducers of Systemic Acquired Resistance. Plant Cell **6:** 1583-1592

Ciolkowski I, Wanke D, Birkenbihl RP, Somssich IE (2008) Studies on DNA-binding selectivity of WRKY transcription factors lend structural clues into WRKY-domain function. Plant Mol Biol **68:** 81-92

Colcombet J and Hirt H (2008) Arabidopsis MAPKs: a complex signalling network involved in multiple biological processes. Biochem. J. **413:** 217–226

Djamei A, Pitzschke A, Nakagami H, Rajh I, Hirt H (2007) Trojan horse strategy in Agrobacterium transformation: abusing MAPK defense signaling. Science **318:** 453-456

Feilner T, Hultschig C, Lee J, Meyer S, Immink RG, Koenig A, Possling A, Seitz H, Beveridge A, Scheel D, Cahill DJ, Lehrach H, Kreutzberger J, Kersten B (2005) High throughput identification of potential Arabidopsis mitogen-activated protein kinases substrates. Mol Cell Proteomics **4:** 1558-1568

Frickey T, Benedito VA, Udvardi M, Weiller G (2008) AffyTrees: facilitating comparative analysis of Affymetrix plant microarray chips. Plant Physiol **146:** 377-386

Fukazawa J, Sakai T, Ishida S, Yamaguchi I, Kamiya Y, Takahashi Y (2000) Repression of shoot growth, a bZIP transcriptional activator, regulates cell elongation by controlling the level of gibberellins. Plant Cell **12:** 901-915

Gao M, Liu J, Bi D, Zhang Z, Cheng F, Chen S, Zhang Y (2008) MEKK1, MKK1/MKK2 and MPK4 function together in a mitogen-activated protein kinase cascade to regulate innate immunity in plants. Cell Res **18:** 1190-1198

Garcia-Hernandez M, Berardini TZ, Chen G, Crist D, Doyle A, Huala E, Knee E, Lambrecht M, Miller N, Mueller LA, Mundodi S, Reiser L, Rhee SY, Scholl R, Tacklind J, Weems DC, Wu Y, Xu I, Yoo D, Yoon J, Zhang P (2002) TAIR: a resource for integrated Arabidopsis data. Funct Integr Genomics **2:** 239-253

Geisler-Lee J, O'Toole N, Ammar R, Provart NJ, Millar AH, Geisler M (2007) A predicted interactome for Arabidopsis. Plant Physiol **145:** 317-329

Heazlewood JL, Verboom RE, Tonti-Filippini J, Small I, Millar AH (2007) SUBA: the Arabidopsis Subcellular Database. Nucleic Acids Res **35:** D213-218

Higo K, Ugawa Y, Iwamoto M, Higo H (1998) PLACE: a database of plant cis-acting regulatory DNA elements. Nucleic Acids Res **26:** 358-359

Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K (2007) WoLF PSORT: protein localization predictor. Nucleic Acids Res **35:** W585-587

Jakoby M, Weisshaar B, Droge-Laser W, Vicente-Carbajosa J, Tiedemann J, Kroj T, Parcy F (2002) bZIP transcription factors in Arabidopsis. Trends Plant Sci **7:** 106-111

Kankainen M, Holm L (2004) POBO, transcription factor binding site verification with bootstrapping. Nucleic Acids Res **32:** W222-229

Kant P, Gordon M, Kant S, Zolla G, Davydov O, Heimer YM, Chalifa-Caspi V, Shaked R, Barak S (2008) Functional-genomics-based identification of genes that regulate Arabidopsis responses to multiple abiotic stresses. Plant Cell Environ **31:** 697-714

Lake JA, Wade RN (2009) Plant-pathogen interactions and elevated CO2: morphological changes in favour of pathogens. J Exp Bot **60:** 3123-3131

Leonhardt N, Kwak JM, Robert N, Waner D, Leonhardt G, Schroeder JI (2004) Microarray expression analyses of Arabidopsis guard cells and isolation of a recessive abscisic acid hypersensitive protein phosphatase 2C mutant. Plant Cell **16:** 596-615

Lescot M, Dehais P, Thijs G, Marchal K, Moreau Y, Van de Peer Y, Rouze P, Rombauts S (2002) PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res **30:** 325-327

Leon P, Sheen J (2003) Sugar and hormone connections. Trends Plant Sci **8:** 110-116

Li J, Krichevsky A, Vaidya M, Tzfira T, Citovsky V (2005) Uncoupling of the functions of the Arabidopsis VIP1 protein in transient and stable plant genetic transformation by Agrobacterium. Proc Natl Acad Sci U S A **102:** 5733-5738

Ma S, Bohnert HJ (2007) Integration of Arabidopsis thaliana stress-related transcript profiles, promoter structures, and cell-specific expression. Genome Biol **8:** R49

Mahony S, Benos PV (2007) STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic Acids Res **35:** W253-258

Nielsen HB, Mundy J, Willenbrock H (2007) Functional Associations by Response Overlap (FARO), a functional genomics approach matching gene expression phenotypes. PLoS One **2:** e676

Obayashi T, Kinoshita K, Nakai K, Shibaoka M, Hayashi S, Saeki M, Shibata D, Saito K, Ohta H (2007) ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. Nucleic Acids Res **35:** D863-869

Petersen K, Fiil BK, Mundy J, Petersen M (2008) Downstream targets of WRKY33. Plant Signal Behav **3:** 1033-1034

Pitzschke A, Djamei A, Bitton F, Hirt H (2009b) A Major Role of the MEKK1-MKK1/2-MPK4 Pathway in ROS Signalling. Mol Plant **2:** 120-137

Pitzschke A, Djamei A, Teige M, Hirt H (2009c) VIP1 response elements mediate mitogen-activated protein kinase 3-induced stress gene expression. Proc Natl Acad Sci U S A **106:** 18414-18419

Pitzschke A, Schikora A, Hirt H (2009a) MAPK cascade signalling networks in plant defence. Curr Opin Plant Biol **12:** 421-426

Popescu SC, Popescu GV, Bachan S, Zhang Z, Gerstein M, Snyder M, Dinesh-Kumar SP (2009) MAPK target networks in Arabidopsis thaliana revealed using functional protein microarrays. Genes Dev **23:** 80-92

Qiu JL, Fiil BK, Petersen K, Nielsen HB, Botanga CJ, Thorgrimsen S, Palma K, Suarez-Rodriguez MC, Sandbech-Clausen S, Lichota J, Brodersen P, Grasser KD, Mattsson O, Glazebrook J, Mundy J, Petersen M (2008b) Arabidopsis MAP kinase 4 regulates gene expression through transcription factor release in the nucleus. Embo J **27:** 2214-2221

Qiu JL, Zhou L, Yun BW, Nielsen HB, Fiil BK, Petersen K, Mackinlay J, Loake GJ, Mundy J, Morris PC (2008a) Arabidopsis mitogen-activated protein kinase kinases MKK1 and MKK2 have overlapping functions in defense signaling mediated by MEKK1, MPK4, and MKS1. Plant Physiol **148:** 212-222

Ringli C, Keller B (1998) Specific interaction of the tomato bZIP transcription factor VSF-1 with a non-palindromic DNA sequence that controls vascular gene expression. Plant Mol Biol **37:** 977-988

Rook F, Corke F, Card R, Munz G, Smith C, Bevan MW (2001) Impaired sucrose-induction mutants reveal the modulation of sugar-induced starch biosynthetic gene expression by abscisic acid signalling. Plant J **26:** 421-433

Rushton PJ, Reinstadler A, Lipka V, Lippok B, Somssich IE (2002) Synthetic plant promoters containing defined regulatory elements provide novel insights into pathogen- and wound-induced signaling. Plant Cell **14:** 749-762

Sato M, Mitra RM, Coller J, Wang D, Spivey NW, Dewdney J, Denoux C, Glazebrook J, Katagiri F (2007) A high-performance, small-scale microarray for expression profiling of many samples in Arabidopsis-pathogen studies. Plant J **49:** 565-577

Spoel SH, Mou Z, Tada Y, Spivey NW, Genschik P, Dong X (2009) Proteasome-mediated turnover of the transcription coactivator NPR1 plays dual roles in regulating plant immunity. Cell **137:** 860-872

Swindell WR (2006) The association among gene expression responses to nine abiotic stress treatments in Arabidopsis thaliana. Genetics **174:** 1811-1824

Thimm O, Blasing O, Gibon Y, Nagel A, Meyer S, Kruger P, Selbig J, Muller LA, Rhee SY, Stitt M (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J **37:** 914-939

Ulker B, Somssich IE (2004) WRKY transcription factors: from DNA binding towards biological function. Curr Opin Plant Biol **7:** 491-498

Urbanczyk-Wochniak E, Usadel B, Thimm O, Nunes-Nesi A, Carrari F, Davy M, Blasing O, Kowalczyk M, Weicht D, Polinceusz A, Meyer S, Stitt M, Fernie AR (2006) Conversion of MapMan to allow the analysis of transcript data from Solanaceous species: effects of genetic and environmental alterations in energy metabolism in the leaf. Plant Mol Biol **60:** 773-792

Usadel B, Poree F, Nagel A, Lohse M, Czedik-Eysenberg A, Stitt M (2009) A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, Maize. Plant Cell Environ **32:** 1211-1229

Wang D, Amornsiripanitch N, Dong X (2006) A genomic approach to identify regulatory nodes in the transcriptional network of systemic acquired resistance in plants. PLoS Pathog **2:** e123

Wang H, Ngwenyama N, Liu Y, Walker JC, Zhang S (2007) Stomatal development and patterning are regulated by environmentally responsive mitogen-activated protein kinases in Arabidopsis. Plant Cell **19:** 63-73

Wang J, Shi H, Mao X, Runzhi L (2006) [Transcription factors networks and their roles in plant responses to environmental stress]. Ying Yong Sheng Tai Xue Bao **17:** 1740-1746

Wang L, Tsuda K, Sato M, Cohen JD, Katagiri F, Glazebrook J (2009) Arabidopsis CaM binding protein CBP60g contributes to MAMP-induced SA accumulation and is involved in disease resistance against Pseudomonas syringae. PLoS Pathog **5:** e1000301

Yan T, Yoo D, Berardini TZ, Mueller LA, Weems DC, Weng S, Cherry JM, Rhee SY (2005) PatMatch: a program for finding patterns in peptide and nucleotide sequences. Nucleic Acids Res **33:** W262-266

Yi H, Richards EJ (2007) A cluster of disease resistance genes in Arabidopsis is coordinately regulated by transcriptional activation and RNA silencing. Plant Cell **19:** 2929-2939

Yuan J, Zhu M, Lightfoot DA, Iqbal MJ, Yang JY, Meksem K (2008) In silico comparison of transcript abundances during Arabidopsis thaliana and Glycine max resistance to Fusarium virguliforme. BMC Genomics **9 Suppl 2:** S6

Zheng Z, Qamar SA, Chen Z, Mengiste T (2006) Arabidopsis WRKY33 transcription factor is required for resistance to necrotrophic fungal pathogens. Plant J **48:** 592-605

Zimmermann P, Hirsch-Hoffmann M, Hennig L, Gruissem W (2004) GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. Plant Physiol **136:** 2621-2632

Figure 1
Flow chart of a possible strategy to identify the DNA motif(s) and promoters targeted by a transcription factor (TF) of interest. Bioinformatic tools are shown in italics.

Signalling Pathway

↓

Protein Kinase/Phosphatase

↓

Top candidate transcription factors

↓ ***Random DNA selection assay***

DNA fragments with candidate motifs

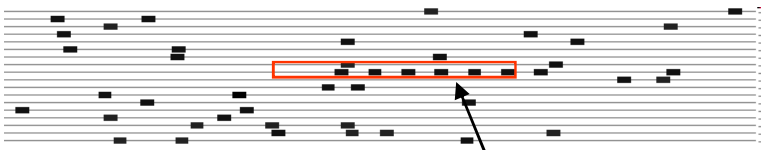
↓ ***STAMP tool***

Experimental validation of TF-motif binding *in vitro*

Identification of candidate target promoters

Hits found:	2344
Sequences with hits:	2218
Sequences searched:	33518
Bytes searched:	16758254
Pattern:	acgagmnnt
Dataset searched:	TAIR9 Loci Upstream Sequences -- 500 bp (DNA)
Download all matches as a textfile	download

TAIR patmatch geneID and promoter download



Visualisation of motif position of promoter (cis-motif seeker tool)

Next Results »

Hit#	Sequence name	# of hits	Hit pattern	Matching Positions		Hit sequence
				start	end	
			AACTCTCGT	60	52	sequence

Filter: candidate genes e.g. with regularly-spaced motifs

↓

Experimental validation *in vivo*

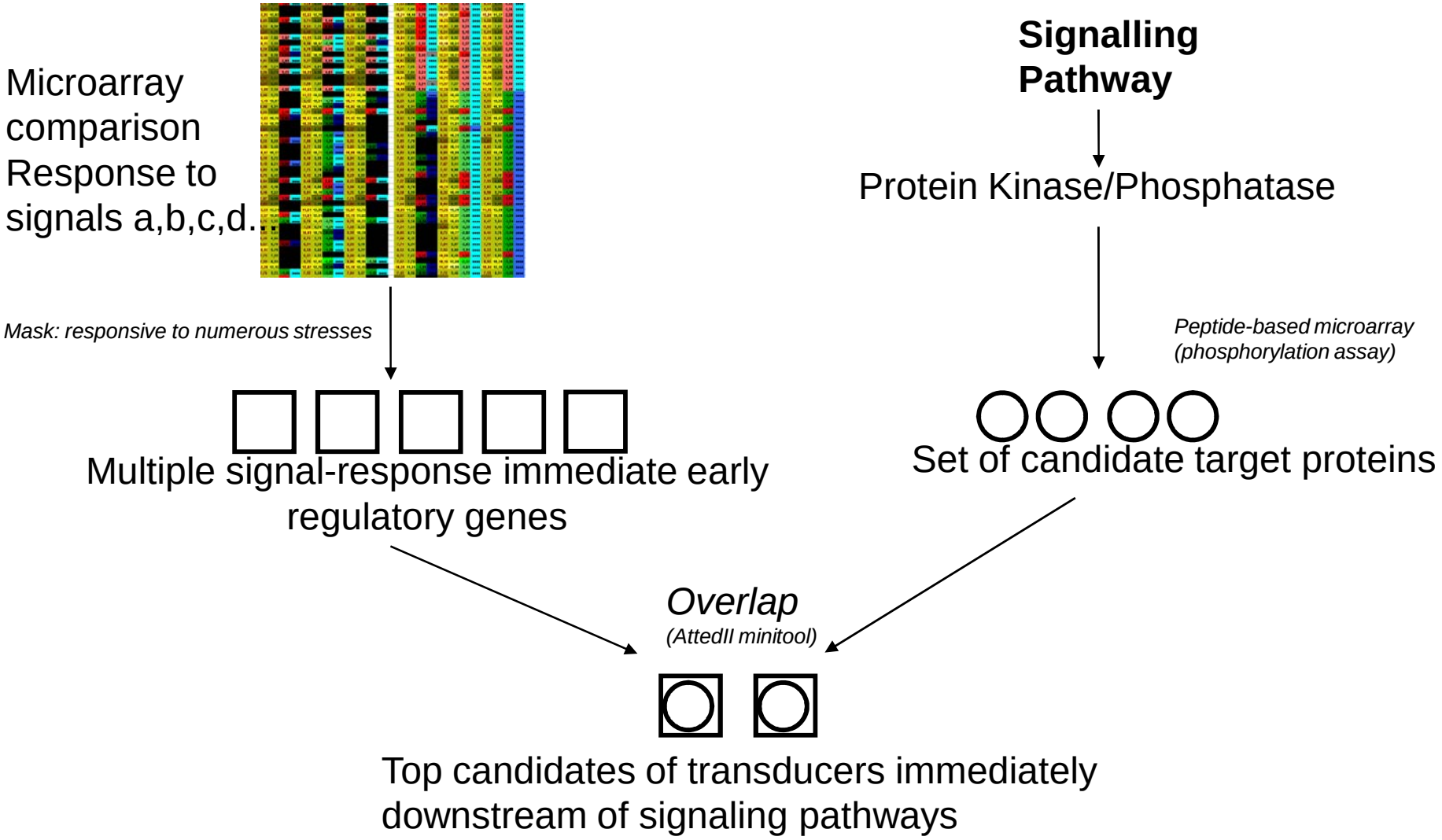


Figure2
Prediction of signalling components through integration of transcriptome and proteome arrays. By searching for the overlap between candidate proteins identified from peptide-based microarrays (e.g. interactors of a regulatory protein in a process of interest; right) and proteins encoded by genes differentially expressed under conditions of interest (left), high-priority candidates involved in the immediate downstream signalling can be defined. As exemplified here, multiple-stress-responsive genes that encode for proteins for which phosphorylation by a stress-associated kinase/phosphatase has been observed, are likely to be key components in early stress signal transduction.