Expression profiles of metabolic models to predict compartmentation of enzymes in multi-compartmental systems

Achuthanunni Chokkathukalam, Mark Poolman, Chiara Ferrazzi and David Fell

cbaunni@brookes.ac.uk

Abstract: Enzymes and other proteins coded by nuclear genes are targeted towards various compartments in the plant cell. Here, we describe a method by which localisation of enzymes in a plant cell may be predicted based on their transcription profile in conjunction with analysis of the structure of the metabolic network. This method uses reaction correlation coefficients to identify reactions in a metabolic model that carry similar flux.

First a correlation matrix for the expression of genes of interest is calculated and the columns clustered hierarchically using the correlation coefficient. The rows clustered using reaction correlation coefficients. In the resulting matrix, we show that the genes in a particular compartment are clustered together and compartmental predictions, with respect to a reference gene can be readily made.

1 Introduction

Spatial organisation of metabolism and other cellular functions is a well known feature of plant cells. Enzymes and other proteins coded by nuclear genes are targeted towards various compartments in the plant cell with the help of the targeting information within their amino acid sequence. Identifying the localisation of proteins is thus an important step towards a broader understanding of the cellular function as a whole and may help in determining the role of thousands of uncharacterised proteins predicted by the genome sequencing projects. Modern organelle-focused experimental approaches can identify proteins in a given compartment. However, reliable protein localisation requires that the technique used must be able to distinguish between genuine organelle residents and contaminating proteins [DDWL04]. Although reasonably pure preparations of some organelles can be achieved, there are many difficulties associated with measuring and characterising proteins that are in a compartment [DHS$^+$06]. Nevertheless, a variety of experimental methods are currently being used to identify protein localisation. Recently chimeric fusion proteins (FPs) and mass spectrometry (MS) techniques have been successfully employed to deduce the localisation of approximately 1100 and 2600 proteins, respectively [HVTF$^+$06]. Although these techniques have accelerated the flow of protein localisation information, the subcellular location of the majority of proteins in a plant cell is still not known.

A relatively simple, low-cost and rapid means to tackle this issue is to employ bioinformatic targeting algorithms to predict protein localisation from amino acid sequence. A number of software tools exists, including TargetP [EBvHN07], Predotar [SPLL04], iPSORT [BTM$^+$02], SubLoc [HS01], MitoProt II [CV96], MITOPRED [GFS04], PeroxiP [EEvHC03], and WoLF PSORT [HPO$^+$07], which can predict proteins targeted towards plastid, cytosol, nucleus, mitochondria, peroxisome or the endoplasmic reticulum. However, the output of such programs has been found to be somewhat inconsistent with each other, or with experimentally determined results [HVTFM05], making them unreliable for some analyses.

The advent of whole-system approaches such as microarrays and metabolomics and the accumulation of such high-throughput data have created new opportunities for studying how reactions are coordinated to meet cellular demands. Microarray experiments monitor the expression of thousands of genes simultaneously. Grouping together genes of similar expression pattern is a general starting point in the analysis of expression data. Similarity between genes is measured by the correlation of their expression profiles and hierarchical clustering methods are used to partition data into clusters of genes exhibiting similar expression patters [IBB04]. Numerous studies have shown that co-expression patterns of gene expression across many microarray datasets form modules of genes that are functionally correlated [WPM$^+$06, MDO$^+$08]. Recently this approach was successfully employed in identifying new genes involved in cellulose synthesis in plants [PWM$^+$05].

Here, we describe a method by which localisation of enzymes may be predicted based on the co-expression profiles of genes coding for reactions in a structural model of plant carbon metabolism. Structural models contain stoichiometries of reactions in a metabolic system. Based on the correlation between these reactions, it can be represented hierarchically as a metabolic tree in which the root node represents the complete system, leaf nodes represent individual reactions, and the intermediate nodes represent metabolic modules capable of the net interconversion of metabolites common to reactions inside and outside the module [PSPF07]. Our technique uses reaction correlation profiles generated from metabolic models together with expression correlation profiles obtained from the microarray data to identify the distribution of enzymes in a particular compartment with respect to the experimentally determined location of a protein representing that compartment.

2 Materials and methods

2.1 Construction of the model of plant carbon metabolism

A structural model of plant carbon metabolism including plastid and cytosol compartments was constructed (Figure 1). The model contains reactions of the Calvin cycle, light reactions and glycolysis and is based, in part, on previous models of plant metabolism constructed in our group [PFR03, Ass05]. Protons, CO_2, pyruvate and sucrose were made external (metabolites that are in constant exchange with the extracellular environment) yielding a model with a total of 53 reactions and 49 metabolites. Reversibility of the reactions was determined based on literature. All modelling and model analysis were performed

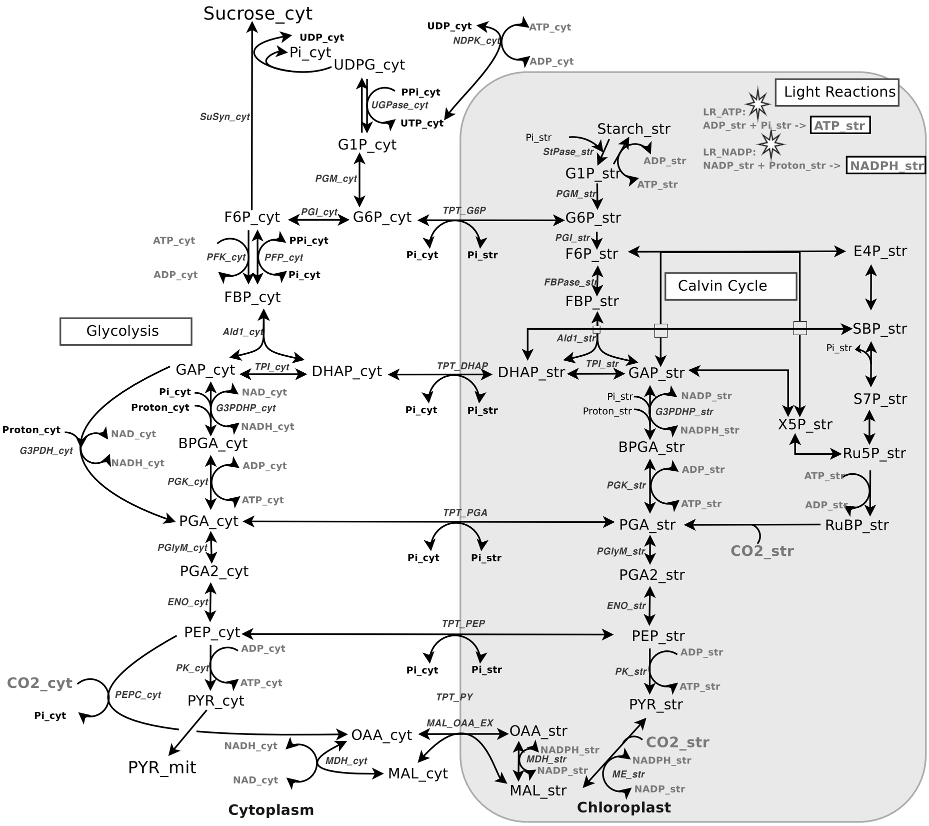

Figure 1: Reaction schema of the model of plant carbon metabolism. For simplicity, the light reactions are depicted here as two separate reactions producing ATP and NADPH. Protons, CO_2 and sucrose are considered external. '_str' and '_cyt' represent the compartments stroma and cytosol, respectively. Notice the transporters connecting reactions of the plastid and the cytosol.

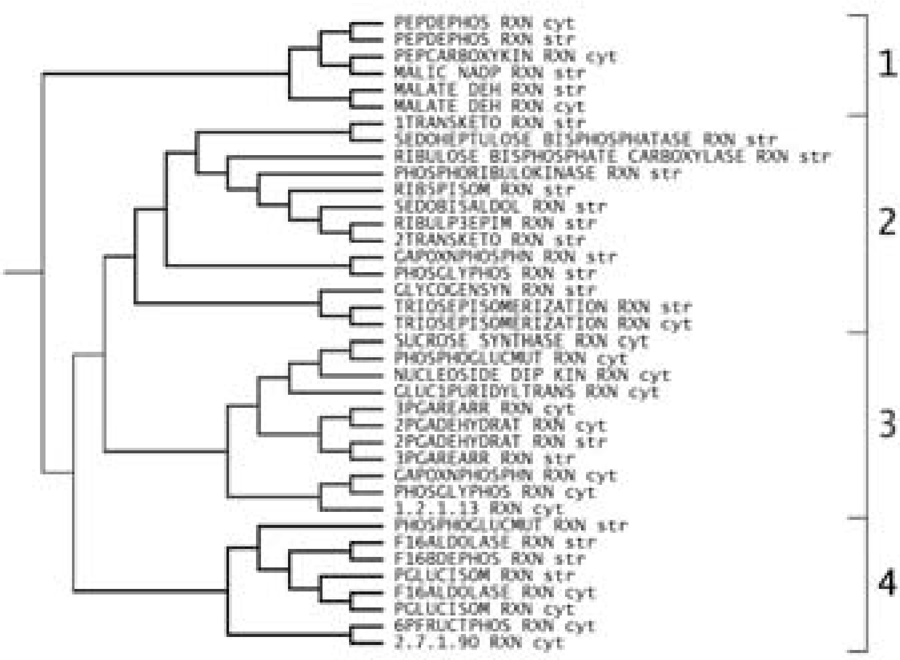

Figure 2: Metabolic tree constructed from the model showing four separate clusters containing reactions capable of net interconversion of metabolites; A. Reactions of the Malate/Oxaloacetate shuttle. B. Calvin cycle reactions. C. Reactions of glycolysis. D. Reactions involved in the regeneration of cytosolic UDP. '_str' and '_cyt' represent the compartments chloroplast and cytosol, respectively.

using the metabolic modelling tool ScrumPy (http://mudshark.brookes.ac.uk) [Poo06].

The model represents the formation of sucrose and pyruvate from the Calvin cycle intermediates transported to the cytosol via specific transport proteins. It contains several reactions such as phosphoglyceromutase, enolase, pyruvate kinase and malate dehydrogenase that are active in both the chloroplast and cytosol. Presence of these reactions in the model will enable us not only to identify their distribution between the compartments but also to distinguish isoforms of genes that code for same reactions in both the compartments. This model is publically available as SBML or in the ScrumPy '.spy' format (http://mudshark.brookes.ac.uk/index.php/User:Cbaunni).

2.2 Expression data analysis of genes coding for reactions in the model

The gene to reaction associations describe the dependence of reactions on genes. The gene to reaction associations in the model were mapped using the AraCyc [ZFT$^+$05] database (http://www.arabidopsis.org/biocyc/index.jsp). The result is a set of genes that potentially code for all the reactions in the model.

The expression data for analysing these genes were obtained from the Nottingham *Arabidopsis* Stock Centre's (NASC) microarray database (http://affymetrix.arabidopsis.info/). The 'super bulk gene' file containing nearly 3500 hybridisations, each with expression level measurements for over 22000 genes represented on the ATH1 array was downloaded (http://affymetrix.arabidopsis.info/narrays/help/usefulfiles.html, March 2009). Expression data from individual experiments were log-transformed; no further modification or scaling was made on the data unless otherwise specified. All microarray data analysis was performed using custom modules designed for ScrumPy.

Expression data for genes ultimately coding for reactions in the model were extracted and a large-scale correlation analysis of expression values between these genes were performed essentially as described by Causton *et al.* [CQB03] by calculating the Pearson's correlation coefficient.

2.3 Clustering and analysis of the correlation matrix

A metabolic tree was generated from the model using the method described in [PSPF07] (Figure 2). The order of the reactions in this tree was used to sort the genes along the rows of the correlation matrix.

The columns of the matrix were hierarchically clustered based on the Pearson's correlation coefficient and an expression correlation tree was generated (Figure 3). Leaves of this tree represent genes in the model and the intermediate nodes are clusters that represent genes sharing similar functions. The columns of the correlation matrix were then sorted in the order of the leaves of the expression correlation tree.

The correlation matrix was imported into TM4-MeV (http://www.tm4.org/mev.html) for visualisation as heatmap [ESBB98]. The metabolic trees were visualised using MEGA phylogenetic tree editor (http://www.megasoftware.net/) [KNDT08].

3 Results and Discussion

3.1 Identification of correlated genes sharing similar flux

Metabolic tree generated from the model contain four separate clusters, each representing reactions capable of net interconversion of metabolites (Figure 2). It is notable that reactions of the Calvin cycle and glycolysis are represented as separate nodes on the tree. Clustering the rows of the correlation matrix based on the genes coding for reactions represented in these nodes can rearrange the heatmap vertically based on the similarities in flux. On the other hand, hierarchically clustering the columns of the correlation matrix grouped genes horizontally depending on their levels of expression. Doing so resulted in the formation of clusters in the heatmap representing genes that are expressed together and code for enzymes that share a similar flux (Figure 4).

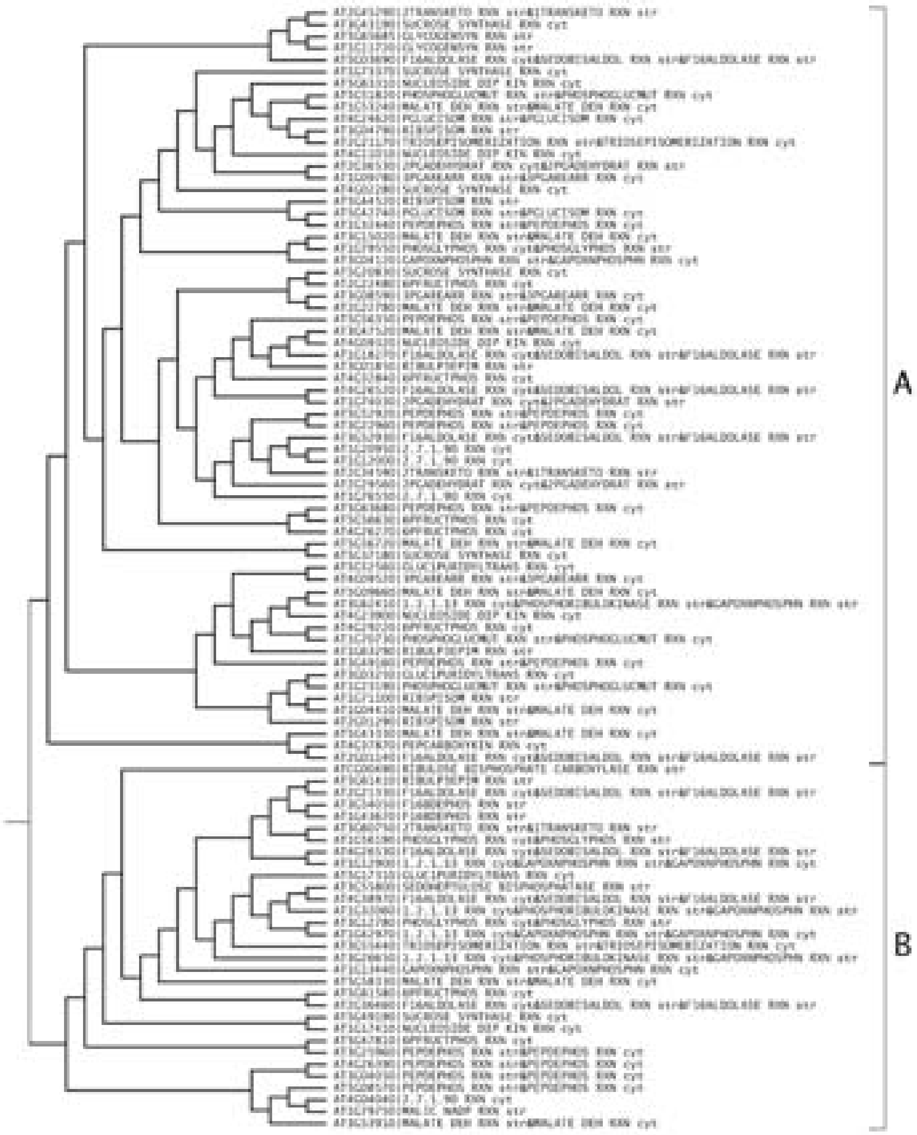

Figure 3: Expression correlation tree generated by hierarchically clustering correlation coefficients of genes coding for reactions in the model showing two separate clusters. A. Genes that predominantly code for reactions in the cytosol correlate with each other B. Genes coding for Calvin cycle intermediates cluster together. '_' is used to separate genes from reactions and '&' is used to distinguish reactions that the gene code for. '_str' and '_cyt' represent the compartments chloroplast and cytosol, respectively.

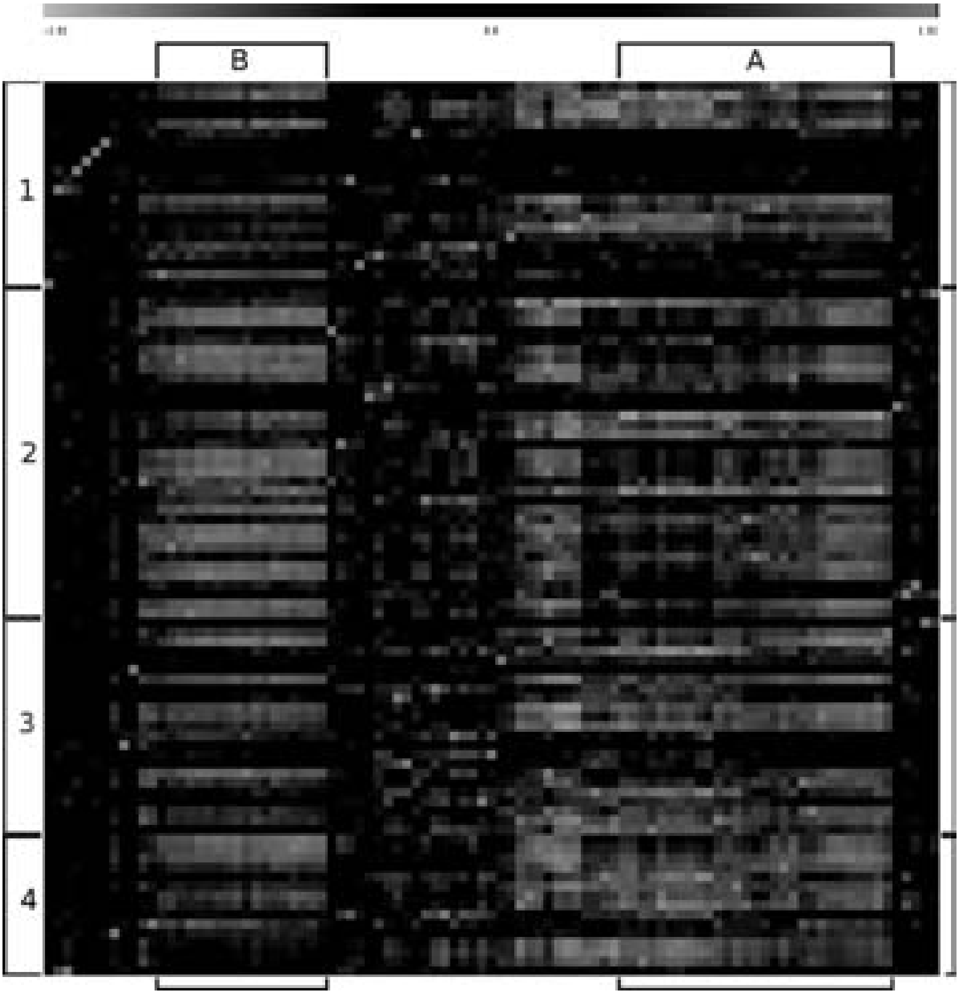

Figure 4: Correlation matrix generated from the expression values of genes coding for reactions in the steady state model. The correlation coefficient ranges from -1 (green) for perfect anticorrelation to +1 (red) for perfect correlation, with zero (black) indicating no relationship. Columns were sorted based on the clustering expression correlation coefficient and rows sorted by clustering based on reaction correlation coefficient. ‘A’ and ‘B’ represent two distinct clusters observed in the correlation matrix (Figure 3). Correlated genes in cluster ‘A’ were found to be highly correlated with reference genes known to be localised in the chloroplast. Whereas correlated genes in cluster ‘B’ showed higher correlation with genes localised in the cytoplasm. 1, 2, 3 and 4 represent clusters in the metabolic tree representing reactions capable of net interconversion of metabolites (Figure 2).

We found that genes coding for reactions in the Calvin cycle are found to be tightly correlated between each other and they cluster together. The same holds true for genes coding for glycolysis reactions. Isoforms of some Calvin cycle genes anticorrelate with other genes coding for reactions of the Calvin cycle. However, those genes that were anticorrelated with the genes of Calvin cycle reactions are found to be tightly correlated with genes of the glycolysis reactions, and vice versa. Similar cases can also be observed in case of the isoforms of glycolytic genes.

A previous study on the transcriptional coordination of metabolic network in *Arabidopsis* suggested that genes coding for reactions in a pathway show tighter levels of correlation [WPM$^+$06]. Results from our study correlates with the above observation and also suggests that the expression profiles of genes can be used to distinguish their compartmentation.

3.2 Identifying compartmentation of genes

Though, this technique is efficient in clustering genes based on their compartmentation, identification of the compartment itself requires a reference gene whose localisation is already known. For example, the plastidic ribulose biphosphate carboxylase (Rubisco) gene ATCG00490 was used as the reference to identify genes localised in the chloroplast. Compartments are identified by filtering out genes that are highly correlated with the reference gene.

The results were compared with the various bioinformatic tools described in Section 1. Comparison with predictions made by bioinformatic tools as a whole was not possible as many of these tools were directed towards particular compartments. Compartmentation of genes that were predicted to be in the chloroplast showed good agreement with tools such as TargetP and Predotar, whereas mitochondrial predictions correlated with MITOPRED and MitoProt II predictions.

This approach was used to predict the localisation of the complete set of genes coding for the reactions in a model containing reactions of the chloroplast, cytosol and mitochondria. Given a good quality microarray expression data containing sufficient experiments that allow reliable statistical analysis, this technique can be used more generically. With the large number of publically available metabolic networks and expression data, this approach may significantly contribute to the identification of enzyme localisation in many different eukaryotic systems.

References

[Ass05] H. Assmus. *Modelling Carbohydrate Metabolism in Potato Tuber Cell.* PhD thesis, Oxford Brookes University, 2005.

[BTM$^+$02] H. Bannai, Y. Tamada, O. Maruyama, K. Nakai, and S. Miyano. Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, 18(2):298–305, 2002.

[CQB03] Helen C. Causton, John Quackenbush, and Alvis Brazma. *Microarray gene expression data analysis: a beginner's guide*. Wiley-Blackwell, 2003.

[CV96] M.G. Claros and P. Vincens. Computational Method to Predict Mitochondrially Imported Proteins and their Targeting Sequences. *European Journal of Biochemistry*, 241:779–786, 1996.

[DDWL04] T.P.J. Dunkley, P. Dupree, R.B. Watson, and K.S. Lilley. The use of isotope-coded affinity tags (ICAT) to study organelle proteomes in Arabidopsis thaliana. *Biochem. Soc. Trans.*, 32(3):520–523, 2004.

[DHS$^+$06] T.P.J Dunkley, S. Hester, I.P. Shadforth, J. Runions, T. Weimar, S.L. Hanton, J.L. Griffin, C. Bessant, F. Brandizzi, C. Hawes, R.B Watson, P. Dupree, and K.S. Lilley. Mapping the Arabidopsis organelle proteome. *Proc. Natl. Acac. Sci. USA*, 103(17):6518–6523, 2006.

[EBvHN07] O. Emanuelsson, S. Brunak, G. von Heijne, and H. Nielsen. Locating proteins in the cell using TargetP, SignalP, and related tools. *Nature Protocols*, 2:953–971, 2007.

[EEvHC03] O. Emanuelsson, A. Elofsson, G. von Heijne, and S. Cristobal. In Silico Prediction of the Peroxisomal Proteome in Fungi, Plants and Animals. *Journal of Molecular Biology*, 330:443–456, 2003.

[ESBB98] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868., 1998.

[GFS04] C. Guda, E. Fahy, and S. Subramaniam. MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins. *Bioinformatics*, 20(11):1785–1794, 2004.

[HPO$^+$07] P. Horton, K-J. Park, T. Obayashi, N. Fujita, H. Harada, C.J. Adams-Collier, and K. Nakai. WoLF PSORT: Protein Localization Predictor. *Nucleic Acids Research*, pages 1–3, 2007.

[HS01] S. Hua and Z. Sun. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 17(8):721–728, 2001.

[HVTF$^+$06] J.L. Heazlewood, R.E. Verboom, J. Tonti-Filippini, I. Small, and A.H. Millar. SUBA: The Arabidopsis Subcellular Database. *Nucleic Acids Research*, 00:1–6, 2006.

[HVTFM05] J.L. Heazlewood, R.E. Verboom, J. Tonti-Filippini, and A.H. Millar. Combining experimental and predicted datasets for determination of the subcellular location of proteins in Arabidopsis. *Plant Physiol.*, 139(2):598–609, 2005.

[IBB04] J. Ihmels, S. Bergmann, and N. Barkai. Defining transcription modules using large-scale gene expression data. *Bioinformatics*, 20(13):1993–2003, 2004.

[KNDT08] S. Kumar, M. Nei, J. Dudley, and K. Tamura. MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. *Briefings in Bioinformatics*, 9(4):299–306, 2008.

[MDO$^+$08] M. Menges, R. Doczi, L. Okresz, P. Morandini, L. Mizzi, M. Soloviev, J.A.H. Murray, and L. Bogre. Comprehensive gene expression atlas for the Arabidopsis MAP kinase signalling pathways. *New Phytologist*, 179(3):643–662, 2008.

[PFR03] M.G. Poolman, D.A. Fell, and C.A. Raines. Elementary modes analysis of photosynthate metabolism in the chloroplast stroma. *Eur. J. Biochem*, 270:430–439, 2003.

[Poo06] M.G. Poolman. ScrumPy - metabolic modelling with Python. *IEE Proceedings Systems Biology*, 153(5):375–378, 2006.

[PSPF07] M.G. Poolman, C. Sebu, M.K. Pidcock, and D.A. Fell. Modular decomposition of metabolic systems via null-space analysis. *Journal of Theoretical Biology*, 249(4):691–705, 2007.

[PWM$^+$05] S. Persson, H. Wei, J. Milne, G.P. Page, and C.R. Somerville. Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proc Natl Acad Sci USA*, 102:8633–8638, 2005.

[SPLL04] I. Small, N. Peeters, F. Legeai, and C. Lurin. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics*, 4(6):1581–90, 2004.

[WPM$^+$06] H. Wei, S. Persson, T. Mehta, V. Srinivasasainagendra, L. Chen, G.P. Page, C. Somerville, and A. Loraine. Transcriptional Coordination of the Metabolic Network in Arabidopsis. *Plant Physiol.*, 142(2):762–774, 2006.

[ZFT$^+$05] P. Zhang, H. Foerster, C.P. Tissier, L. Mueller, S. Paley, P.D. Karp, and S.Y. Rhee. MetaCyc and AraCyc. Metabolic pathway databases for plant research. *Plant Physiol.*, 138:27–37, 2005.