D954–D958 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 11 October 2007*
doi:10.1093/nar/gkm835

MetaCrop: a detailed database of crop plant metabolism

Eva Grafahrend-Belau, Stephan Weise, Dirk Koschützki, Uwe Scholz, Björn H. Junker and Falk Schreiber*

Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstrasse 3, D-06466 Gatersleben, Germany

Received August 15, 2007; Revised and Accepted September 21, 2007

ABSTRACT

MetaCrop is a manually curated repository of high quality information concerning the metabolism of crop plants. This includes pathway diagrams, reactions, locations, transport processes, reaction kinetics, taxonomy and literature. MetaCrop provides detailed information on six major crop plants with high agronomical importance and initial information about several other plants. The web interface supports an easy exploration of the information from overview pathways to single reactions and therefore helps users to understand the metabolism of crop plants. It also allows model creation and automatic data export for detailed models of metabolic pathways therefore supporting systems biology approaches. The MetaCrop database is accessible at http://metacrop.ipk-gatersleben.de.

INTRODUCTION

Crop plants are the major source of human nutrition and important contributors to chemical feedstocks and renewable fuels (1–3). An in-depth understanding of the plant's metabolism is helpful for the improvement of their growth and yield (4,5). Data requirements in metabolic research are quite diverse: while some experts are interested in a qualitative global view of metabolism, others need detailed information about single reactions. Additionally, researchers investigating metabolism often have to rely on databases with unclear data quality resulting from genome-based metabolic network predictions. The situation in crop plant research is furthermore complicated by the fact that only one crop plant (*Oryza sativa*, rice) has been sequenced so far (6,7). An example that requires detailed metabolic information is the generation of models to quantitatively simulate complex biochemical networks, an area which is of increasing interest in systems biology. While repositories for such models exist, the collection of information necessary for model creation remains a time-consuming manual task and only very few models for crop plants exist at all.

Here we present MetaCrop, a database that contains manually curated, highly detailed information about metabolic pathways in crop plants, including location information, transport processes and reaction kinetics. The web interface supports the exploration of the information from overview pathways to single reactions, data export and the creation of detailed models of metabolic pathways. With these features MetaCrop supports crop plant research in several ways: it improves the understanding of the metabolism, especially if one wants to get both a general overview and specific details for selected pathways. It allows the usage of the crop plant specific information in other tools, for example, to investigate experimental data in the network context. And it helps in creating models of metabolic processes for simulation approaches and *in silico* experiments.

DATABASE DESCRIPTION

Content

MetaCrop contains hand-curated information of about 40 major metabolic pathways in various crop plants with special emphasis on the metabolism of agronomically important organs such as seed and tuber. Species of both monocotyledons and dicotyledons are represented. Reactions incorporate information about involved enzymes (e.g. EC and CAS number), metabolites (e.g. CAS number, molecular weight and chemical formula), stoichiometry and detailed location (species, organ, tissue, compartment and developmental stage). Furthermore, for central metabolism (sucrose breakdown, glycolysis, TCA cycle) kinetic data is available for the reactions. References and relevant PubMed IDs are given. In order to have a controlled vocabulary allowing the comparison of data from different sources ontology terms were used (8,9).

Currently the database focuses on the monocotyledon species *Hordeum vulgare* (barley), *Triticum aestivum* (wheat), *Oryza sativa* (rice), *Zea mays* (maize) and the

*To whom correspondence should be addressed. Tel: +49 39482 5753; Fax: +49 39482 5407; Email: schreibe@ipk-gatersleben.de

Table 1. Information contained in MetaCrop

	Hordeum vulgare	*Triticum aestivum*	*Oryza sativa*	*Zea mays*	*Solanum tuberosum*	*Brassica napus*	Total[a]
Pathways	36	33	34	34	34	26	38
Enzymatic reactions	291	271	278	273	207	168	392
Transport processes	7	6	9	27	14	7	59
Compartments	4	4	4	3	3	3	5
References	382	347	340	346	252	204	734

[a]Including other plants; pathways, reactions and other information occurring in more than one plant are only listed once.

dicotyledon species *Solanum tuberosum* (potato) and *Brassica napus* (canola). Additional data of other crop and non-crop plants is currently being added to the database. In total, about 400 enzymatic reactions, 60 transport processes, 5 compartments and 740 references are represented in MetaCrop (see Table 1, content as of July 2007). In order to enable the export of detailed metabolic networks for systems biology approaches, most of the data contained in the database corresponds to biochemical data (e.g. taxon-specific enzymatic information). In the case of missing biochemical information, proteomic information and genetic information, respectively, is represented for a given enzymatic reaction or transport process.

Web interface

The web interface of the database is accessible at http://metacrop.ipk-gatersleben.de. It allows detailed browsing and searching of data, user feedback and data export. Figure 1 shows some screenshots of the MetaCrop web interface starting with a complete pathway (sucrose breakdown in dicotyledon species including compartmentalization, transporters and isoenzymes) to detailed information about reaction kinetics. Additionally to searchable data tables, the user is guided by clickable image maps of the pathways. Entire pathways containing all available information on the respective reactions and metabolites can be downloaded in the standardized systems biology exchange format systems biology markup language (SBML) (10), which can be imported into modelling tools such as COPASI (11,12).

The functionality of the web interface is documented in a tutorial available on the website. It is also possible to edit entries, extend the content of MetaCrop and create user-specific models. To ensure data quality, such changes cannot be done anonymously. Users interested in these functionalities are invited to obtain an editing account for MetaCrop. Changes performed by all accounts are logged and checked by curators to guarantee consistency and quality of the inserted data. The web interface is based on the Oracle Application Express technology.

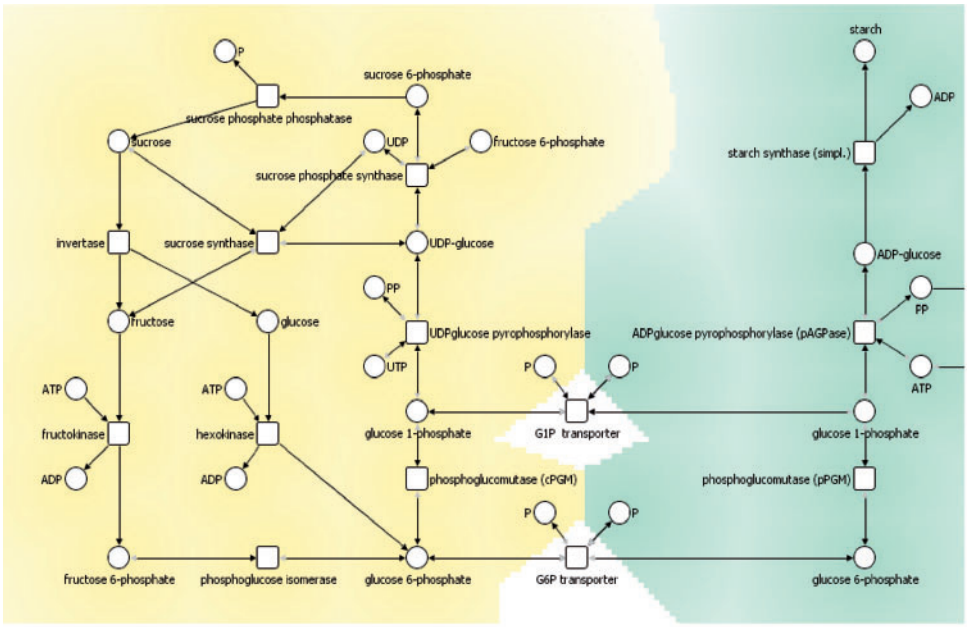
Database implementation

MetaCrop uses the information system Meta-All (13) and is based on the database management system Oracle. The database schema comprises 51 relational tables and can be divided into several parts. The main parts are conversions, substances, pathways, locations, references and versioning. Conversions and substances are the central parts of the schema. A conversion is a reaction or a translocation, which is either active or passive. Substances comprise transporters, enzymes, metabolites and macromolecules. They take place in conversions and play certain roles, such as reactant or product, modulator, catalyst, etc. All necessary information, e.g. name, formula or kinetic data, can be stored together with conversions and substances. In order to distinguish data originating from different publications, each record can be enriched by reference information. The term location describes a combination of taxonomy, developmental stage and cytology of plants in order to distinguish where and when conversions take place. Therefore, controlled vocabulary is used. Additionally, the database schema supports parallel versioning of data records, e.g. in case of different opinions of experimentalists. Finally, pathways are combinations of conversions taking place at a certain location.

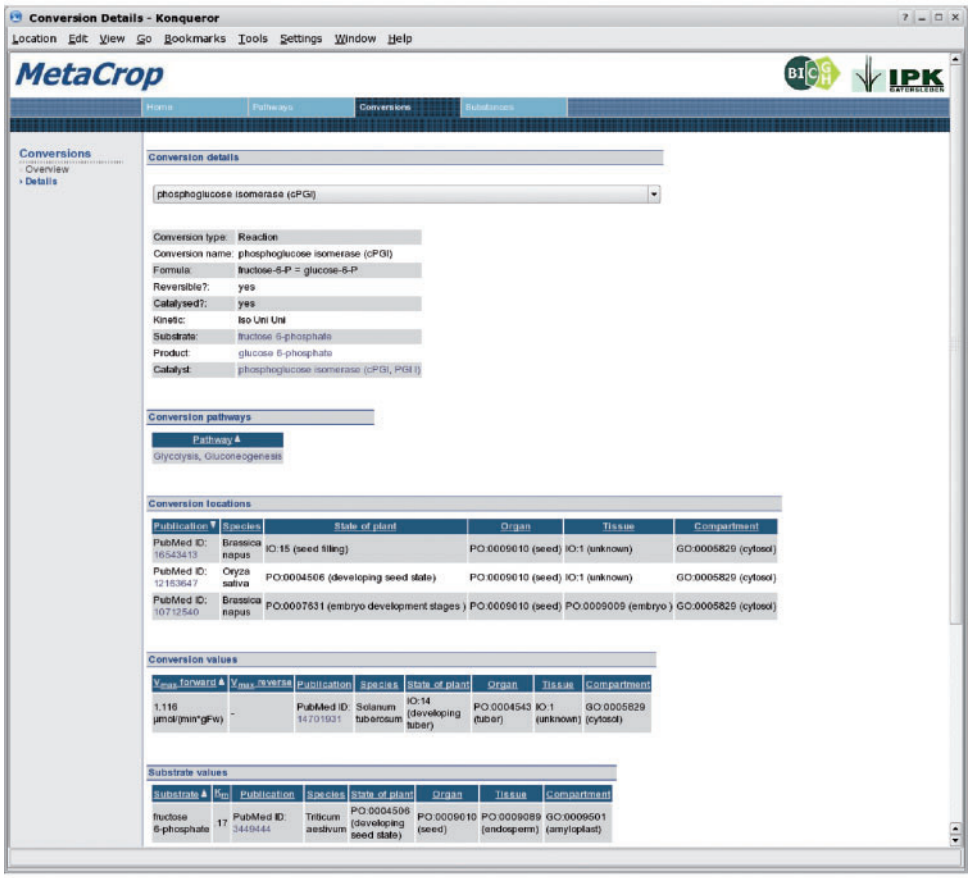
The complete information represented in MetaCrop is also available as a dump of the database, i.e. the data is available for bulk download. The dump can easily be imported into a user's instance of the open source information system Meta-All (13), therefore enabling users to run their local version of the database.

CURATION, QUALITY ASSURANCE, COMPLETENESS AND CONTINUATION

All information was extracted manually through an extensive survey of primary literature and online databases. Literature-based information was derived from about 800 papers of plant biochemical and physiological journals as well as from respective textbooks (e.g. (14,15)). Furthermore, some of the information was manually extracted from online databases providing pathway-related information: KEGG PATHWAY ((16), http://www.genome.jp/kegg/pathway.html), EGENES ((17), http://www.genome.jp/kegg-bin/create_kegg_menu?category=plants_egenes), AraCyc ((18), http://www.arabidopsis.org/biocyc/index.jsp), MetaCyc ((19), http://metacyc.org/), RiceCyc (http://www.gramene.org/pathway/), Reactome ((20), http://www.reactome.org/); enzyme-related information: BRENDA ((21), http://www.brenda-enzymes.info/), ExPASy-ENZYME ((22), http://expasy.org/enzyme/); protein-related information: Swiss-Prot/TrEMBL ((23), http://www.expasy.org/sprot/); metabolite-related information: PubChem (http://pubchem.ncbi), KEGG LIGAND ((16), http://www.genome.jp/kegg/ligand.html);



(a)



(b)

Figure 1. Screenshots of the web interface of MetaCrop. (**a**) A pathway (sucrose breakdown in dicotyledon species, which shows compartmentalization, transporter and isoenzymes); (**b**) Information connected to pathways: conversion details (cytosolic phosphoglucose isomerase): stoichiometry, catalyst, metabolites, conversion location, subset of taxon-specific kinetic parameters (vmax, km) given for cytosolic phosphoglucose isomerase.

transporter-related information: ARAMEMNON ((24), http://aramemnon.botanik.uni-koeln.de/); kinetic information: BRENDA ((21), http://www.brenda-enzymes.info/)).

For quality assurance information inferred from databases has been checked against literature. To enable the trace back of information and further reading, references and corresponding PubMed IDs are given where available. Controlled vocabulary (e.g. ontology terms from Plant Ontology (8) and Gene Ontology (9)) was used to ensure consistency and to allow the comparison of data from different sources. Currently MetaCrop contains most of the pathways of central metabolism in higher plants (e.g. metabolism of carbohydrates, amino acids, lipids, energy, cofactors and nucleotides). With respect to crop plant metabolism, special emphasis is laid on pathways of seed and tuber metabolism such as the sucrose breakdown pathway. While our current focus is on updating pathways with incomplete information, we plan to extend the information stored in MetaCrop to pathways of plant secondary metabolism. The extension of MetaCrop is primary done inhouse; however, registered users can edit entries and extend the content of MetaCrop and therefore may in the future also contribute to the extension of the database.

APPLICATION OF THE METACROP DATABASE

MetaCrop can be used for a wide variety of applications in crop plant research. It helps in understanding the metabolism at different levels of detail, it allows the use of crop plant specific information in other tools for further investigations, and it supports the creation of models of metabolism for simulation approaches. Two example applications are as follows:

Mathematical analysis of metabolic pathways

The in-depth mathematical analysis of a pathway of interest will generally consist of two main steps, which are (i) investigation of the structural properties and capabilities of the pathway with tools such as CellNetAnalyzer (25) and (ii) detailed analysis of the kinetic characteristics of the system with modelling and simulation tools such as COPASI (11). MetaCrop supports these processes at various steps. It contains all necessary information for structural pathway analysis, and for central metabolism also detailed kinetic data for kinetic pathway analysis. Furthermore, the above-mentioned tools are able to read the files exported from MetaCrop in the standardized SBML format (10). Once imported into these tools, the pathways can serve as a starting point for structural or kinetic metabolic models.

Investigation of -omics data in the context of metabolic networks

Network-related analysis of high-throughput data involves the mapping of experimental data onto related pathways and the investigation of this integrated data. Such functionality is provided by tools such as VANTED (26), a system for the visualization and analysis of networks with related experimental data. Data from large-scale biochemical experiments can be uploaded into the software and then mapped on a network that is either drawn with the tool itself or imported, for example, from a SBML file. VANTED enables users to present and analyse transcript, enzyme, proteomics and metabolite data in the context of underlying networks such as metabolic pathways from MetaCrop. Several analysis methods implemented in such software systems help in further investigation of the data.

DISCUSSION

MetaCrop contains comprehensive, original, high-quality data about crop plant metabolism. While most of the existing metabolic pathway databases do not contain any plant-specific information, there exist a few multi-organism databases such as MetaCyc (19), BRENDA (21) and KEGG (16) comprising information about plant metabolism. The transcriptome-based database EGENES (17) is a multi-species plant database, which currently consists of 25 plant species (release 41.0, January 2007). The database integrates plant genomic information (EST contigs) and pathway information (pathway maps derived from KEGG reference pathways), thus offering an overview of fundamental biological processes in plants. In addition to these multi-species databases, there exist a few species-specific crop plant databases such as the pathway/genome databases RiceCyc (http://www.gramene.org/pathway/) and SolCyc (http://www.sgn.cornell.edu/tools/solcyc/). However, most of these single- and multi-species databases only contain little or no hand-curated information due to genome- or EST-based pathway predictions or do not support model creation and model export in SBML. Furthermore, highly specific information such as kinetic data, compartment-specific information or transport processes are often lacking and most of the databases are limited to read-only access not allowing for user-specific interaction, editing and extending.

Similar to MetaCrop the pathway databases AraCyc ((18), http://www.arabidopsis.org/biocyc/index.jsp) and MetNetDB ((27), http://www.metnetdb.org) contain detailed information about plant metabolism. AraCyc is a pathway/genome database that contains enzymes and pathways found in the model plant *Arabidopsis* (*Arabidopsis thaliana*). The Metabolic Networking Data Base (MetNetDB) contains information on metabolic and regulatory networks in *Arabidopsis*, which are derived from a combination of online databases and input from biologists in their area of expertise. Both databases are under continued curation and contain highly specific information such as compartment-specific information or transport processes. However, both databases currently only contain information about the model plant *Arabidopsis*.

CONCLUSION

MetaCrop is an ongoing project and currently consists largely of a collection of manually curated data about six

major crop plants, interactive interaction methods via the web interface and export functionalities. Our vision for the database is in two directions: the further curation of information and the improvement of the web interface. We plan to extend the information stored in MetaCrop to secondary pathways and to include other important crop plants such as *Glycin max* (soybean), *Solanum lycopersicum* (tomato), *Helianthus annuus* (sunflower) and *Secale cereale* (rye). For the web interface work is underway to implement methods to take advantage of the taxonomy and localization information in MetaCrop such that, for example, if information is not available for a specific species it can be derived from information of closely related species.

ACKNOWLEDGEMENTS

This work was partly supported by the German Federal Ministry of Education and Research (grant 0312706A). Funding to pay the Open Access publication charges for this article was provided by the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben.

Conflict of interest statement. None declared.

REFERENCES

1. Grusak,M.A. and DellaPenna,D. (1999) Improving the nutrient composition of plants to enhance human nutrition and health. *Annu. Rev. Plant Physiol. Plant Mol. Biol.*, **50**, 133–161.
2. Metzger,J.O. and Bornscheuer,U. (2006) Lipids as renewable resources: current state of chemical and biotechnological conversion and diversification. *Appl. Microbiol. Biotechnol.*, **71**, 13–22.
3. Tilman,D., Hill,J. and Lehman,C. (2006) Carbon-negative biofuels from low-input high-diversity grassland biomass. *Science*, **314**, 1598–1600.
4. Jenner,H.L. (2003) Transgenesis and yield: what are our targets? *Trends Biotechnol.*, **21**, 190–192.
5. Carrari,F., Urbanczyk-Wochniak,E., Willmitzer,L. and Fernie,A.R. (2003) Engineering central metabolism in crop species: learning the system. *Metab. Eng.*, **5**, 191–200.
6. Yu,J., Hu,S., Wang,J., Wong,G.K.-S., Li,S., Liu,B., Deng,Y., Dai,L., Zhou,Y. *et al.* (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science*, **296**, 79–92.
7. Goff,S.A., Ricke,D., Lan,T.-H., Presting,G., Wang,R., Dunn,M., Glazebrook,J., Sessions,A., Oeller,P. *et al.* (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science*, **296**, 92–100.
8. Jaiswal,P., Avraham,S., Ilic,K., Kellogg,E.A., McCouch,S., Pujar,A., Reiser,L., Rhee,S. Y., Sachs,M.M. *et al.* (2005) Plant Ontology (PO): a controlled vocabulary of plant structures and growth stages. *Comp. Funct. Genomics*, **6**, 388–397.
9. Gene Ontology Consertium (2006) The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.*, **34**(Suppl. 1), D322–D326.
10. Hucka,M., Finney,A., Sauro,H.M., Bolouri,H., Doyle,J.C., Kitano,H., Arkin,A.P., Bornstein,B.J., Bray,D. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–531.
11. Hoops,S., Sahle,S., Gauges,R., Lee,C., Pahle,J., Simus,N., Singhal,M., Xu,L., Mendes,P.U. *et al.* (2006) COPASI—a complex pathway simulator. *Bioinformatics*, **22**, 3067–3074.
12. Alves,R., Antunes,F. and Salvador,A. (2006) Tools for kinetic modeling of biochemical networks. *Nat. Biotechnol.*, **24**, 667–672.
13. Weise,S., Grosse,I., Klukas,C., Koschützki,D., Scholz,U., Schreiber,F. and Junker,B.H. (2006) Meta-All: a system for managing metabolic pathway information. *BMC Bioinformatics*, **7**, e465.
14. Buchanan,B.B., Gruissem,W. and Russel,L.J. (2000) *Biochemistry & Molecular Biology of Plants*. American Society of Plant Physiologists, Rockeville, MD.
15. Bewley,J.D. and Black,M. (1994) *Seeds: Physiology of Development and Germination,* 2nd edn. Plenum Press, New York, USA.
16. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**(Suppl. 1), D354–D357.
17. Masoudi-Nejad,A., Goto,S., Jauregui,R., Ito,M., Kawashima,S., Moriya,Y., Endo,T. and Kanehisa,M. (2007) EGENES: transcriptome-based plant database of genes with metabolic pathway information and expressed sequence tag indices in KEGG. *Plant Physiol.*, **144**, 857–866.
18. Rhee,S.Y., Zhang,P., Foerster,H. and Tissier,C. (2006) AraCyc: overview of an Arabidopsis metabolism database and its applications for plant research. In Saito,K., Dixon,R.A. and Willmitzer,L. (eds), *Plant Metabolomics*. Springer Berlin, Heidelberg, pp. 141–154.
19. Caspi,R., Foerster,H., Fulcher,C., Hopkinson,R., Ingraham,J., Kaipa,P., Krummenacker,M., Paley,S., Pick,J. *et al.* (2006) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.*, **34**(Suppl. 1), D511–D516.
20. Vastrik,I., D'Eustachio,P., Schmidt,E., Joshi-Tope,G., Gopinath,G., Croft,D., de Bono,B., Gillespie,M., Jassal,B. *et al.* (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.*, **8**, R39.
21. Barthelmes,J., Ebeling,C., Chang,A., Schomburg,I. and Schomburg,D. (2007) BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. *Nucleic Acids Res.*, **35**(Suppl. 1), D511–D514.
22. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.
23. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement trEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
24. Schwacke,R., Schneider,A., van der Graaff,E., Fischer,K., Catoni,E., Desimone,M., Frommer,W.B., Flügge,U.-I. and Kunze,R. (2003) ARAMEMNON: a novel database for Arabidopsis integral membrane proteins. *Plant Physiol.*, **131**, 16–26.
25. Klamt,S., Saez-Rodriguez,J. and Gilles,E.D. (2007) Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Syst. Biol.*, **1**, e2.
26. Junker,B.H., Klukas,C. and Schreiber,F. (2006) VANTED: a system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics*, **7**, e109.
27. Wurtele,E.S., Li,J., Diao,L., Zhang,H., Foster,C.M., Fatland,B., Dickerson,J., Brown,A., Cox,Z. *et al.* (2003) MetNet: software to build and model the biogenetic lattice of Arabidopsis. *Comp. Funct. Genomics*, **4**, 239–245.