ICBAA2017-14

BIOLOGICAL PATHWAY RECONSTRUCTION VIA INTEGRATED OMICS DATA ANALYSIS

Zeti-Azura Mohamed-Hussein[1,2] and Loke Kok Keong[2]

[1]*Pusat Pengajian Biosains & Bioteknologi, Fakulti Sains & Teknologi, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor.*
[2]*Institut Biologi Sistem (INBIOSIS), Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor.*
zeti.hussein@ukm.edu.my

Abstract: Systems biology approach has generated numerous amount of heterogeneous data that requires computational analysis for biological discovery. Research has been carried out on a popular Malaysian herb, kesum (*Polygonum minus* Huds.), to investigate the secondary metabolite pathways involved in the production of interesting compounds identified from previous studies. Extensive bioinformatic analysis was carried out on the metabolomic and transcriptomic data of *P. minus* towards the reconstruction of its secondary metabolite pathways. Specific databases were successfully developed to systematically store and manage the heterogeneous data generated from *P. minus* systems biology research. This study has successfully established the first *P. minus* metabolic pathway by mapping 1385 genes and 295 metabolites onto the skeleton pathways obtained from 335 individual KEGG sub maps.

Keywords: pathway reconstruction, data integration, metabolic pathway, omics,

INTRODUCTION

Plants are the greatest source of food, energy and various valuable compounds. The ever-expanding demand for the production of food, feed, medicine, and biofuel from plants has prompted the sequencing of plant omics. The developing field of plant systems biology has provided outstanding insights into how these products are synthesized with an ultimate goal towards in depth understanding of the genotype-phenotype relationship in cellular systems (Kell, 2002; Benfey and Mitchell-Olds, 2008; Weckwerth, 2011). Comprehensive measuring and analysing genes, transcripts, proteins and metabolites are possible with recent technical advances in high-throughput sequencing and various analytical instruments in omics technologies (Fukushima et al., 2009; Lei et al., 2011; Lucas et al., 2011). These technologies provide platforms that can monitor the cellular inventory as well providing the opportunity to evaluate cellular behaviours from a multi-level perspective to enhance our understanding on plant systems (Saito and Matsuda, 2010; Dhondt et al., 2013). Major effective and efficient approaches to analyse omics data are network- and pathway analysis. Metabolic networks reconstructed from omics data can help visualize and analyse large-scale experimental data, predict metabolic phenotypes, discover enzymes, engineer metabolic pathways and study metabolic pathway evolution. Pathway analysis is a knowledge-based approach that involves the associated biochemical pathways hence it can be used to facilitate enzyme discovery and metabolic engineering. In addition, using such systems-level annotations will also enable researchers studying individual genes and mutants to contextualize their findings within the overall metabolic scheme of an organism, thereby providing a framework for assessing the broader roles of their genes of interest.

METHODS

Here we describe our approach in reconstructing a single-species metabolic pathway by mapping omics data (e.g. transcriptomic and metabolomic) of *P. minus*. A standard two-step method in reconstructing a single-species metabolic pathway includes (a) annotation analysis using various bioinformatic tools on transcripts and metabolites, (b) reconstruction of *P. minus* biosynthetic pathway by mapping the annotated omics data into the pathway template obtained from KEGG Pathway database.

RESULTS AND DISCUSSION

P. minus metabolic pathways were reconstructed from the transcriptomic and metabolomic data using Cytoscape. It was then manually curated using a comparative analysis approach, which included comparison of metabolic pathways with other organisms. A generalized scheme for the metabolism-centered approach was used as standards for determining the presence of pathways and enzymes. For each pathway predicted in any species-specific pathway, biochemical evidence in the literature was searched manually to determine if the pathway is present in that particular organism (in this case, it is plants). When a pathway present in plants was not adequately represented in MetaCyc, it was modified on the basis of information from KEGG, the literature and plant biochemistry text books. Three data sources were used intensively as references. "Missing" metabolic proteins for which no gene was identified in a *P. minus* pathway were searched for in the plant genomes and non-redundant protein databases using TBLASTN and BLASTP, respectively. The thresholds used for identification of the *P. minus* ortholog of a plant protein are 80% coverage and 70% identity, which were similar to those used in the Ensembl gene annotation. Additional orthologs were assigned if the best BLAST hit included > 50% and exactly matched > 90% of the query protein sequence. Reactions mediated by those enzymes were also searched for in the literature and BRENDA. The bioinformatic and biochemical evidence used for gene annotation were referenced and documented. Comparative metabolic analysis was performed against MetaCyc, Aracyc, MtruncatulaCyc and BrapafpcCYC followed by manual inspection to identify metabolic differences among these species.

The initial build of P. minus contained 1385 genes and 295 metabolites mapped onto 335 metabolic pathways. 184 metabolic pathways contained one or more pathway holes, which are defined as reactions in which the organism-specific enzyme has not yet been identified. The total number of pathway holes was 56% of the total known reactions in pathways due to many unidentified genes encoding enzymes in known pathways. To improve metabolic reconstruction of *P. minus* we manually reviewed 553 metabolic pathways present in MetaCyc and AraCyc and also predicted in the automated reconstructions for *Medicago truncatula and Brassica napus*. Consequently, the curated *P. minus* reconstructed pathway consists of 230 pathways from the automated reconstruction and 105 pathways that were manually added. Most of these omics data were distributed in the phenylpropanoid biosynthetic pathway. It is one of the many plant secondary metabolite classes that play a very important role in producing complex compounds such as lignin, colouring agents and aromatic compounds. Phenylpropanoids are a diverse group of compounds derived from the carbon skeleton of phenylalanine that are involved in plant defense, structural support, and survival (Vogt 2010). KOBAS analysis showed the distribution omics data in these subpathways such as terpene biosynthesis, flavonoid biosynthesis, limonine and penine degradation, flavonol and flavone biosynthesis and indole alkaloid biosynthesis. These pathways are highly involved in the production of compounds with aromatic and colouring properties.

Table 1: Mapping of *P.* minus transcriptomics data on the pathway template obtained from KEGG.

KEGG Metabolic Pathway	**Number of KEGG Pathway Entity**	**No *P. minus mapping***
KO00983 Drug metabolisme	22	22
KO00471 D-Glutamine and D-glutamate metabolisme	6	6
KO00941 Flavonoid Biosynthesis*	19	17
KO00780 Biotin metabolisme	19	17
KO00903 Limonene and pinene degradation *	19	16
KO00100 Steroid biosynthesis	30	25
KO00944 Flavone dan flavonol biosynthesis*	12	10
KO00120 Primary bile acid biosynthesis	18	15
KO00410 β-Alanine	40	33
KO00363 Bisphenol Degradation	11	9
KO00901 Indole alkaloid biosynthesis *	10	8
KO00473 D-Alanine metabolisme	5	4
KO00531 Glycosaminoglycan degradation	14	11
KO00300 Lysine biosynthesis	46	36
KO00592 α-Linolenic metabolisme	23	18
KO00670 Metabolism of cofactors and vitamins	30	23
KO00908 Zeatin biosynthesis *	8	6
KO00351 DDT degradation	4	3
KO00770 Pantothenate and CoA biosynthesis	34	25
KO00930 Caprolactam	15	11

CONCLUSIONS

P. minus reconstructed pathway consists of subpathways that are abundantly distributed in the phenylpropanoid biosynthetic pathway; *where it* serves as a rich source of metabolites in plants, being required for the biosynthesis of lignin, and serving as a starting point for the production of many other important compounds, such as the flavonoids, coumarins, and lignans. Compounds with these roles, or even with no known roles, are often referred to as “secondary metabolites” because they have no apparent involvement in basal cellular processes. This finding will surely become an open question on why there are many phenylpropanoid subpathways in *P. minus.*

REFERENCES

Benfey, P. N & Mitchell-Olds, T. 2008. From genotype to phenotype: systems biology meets natural variation. Science 320, 495-497.

Kell, D. B. 2002. Genotype-phenotype mapping: genes as computer programs. Trends in Genet. 18, 555-559.

Weckwerth, W. 2011. Green systems biology - from single genomes, proteomues and metabolomes to ecosystems research and biotechnology. J. Proteomics. 75, 284-305.

Saito, K., & Matsuda, F. 2010. Metabolomics for functional genomics, systems biology and biotechnology. Annu. Rev. Plant Biol. 61, 463-489.

Dhondt, S., Wuyts, N. & Inze, D. 2013. Cell to whole plant phenotyping: the best is yet to come. Trends in Plant Sci 18, 428-439.

Fukushima, A., Kusano, M., Redestig, H., Arita, M. & Saito, K. 2009. Integrated omics approach in plant systems biology. Curr. Opin. Chem. Biol. 13, 532-538.

Lucas, M., Laplaze, L. & Benner, M.J. 2011. Plant systems biology: network matters. Plant Cell Environment. 34, 535-553.

Vogt T. 2010. Phenylpropanoid biosynthesis. Mol. Plant. 3, 2–20.