An Integrative Bioinformatics Framework for Genome-scale Multiple Level Network Reconstruction of Rice

Lili Liu[1], Qian Mei[1], Zhenning Yu[1], Tianhao Sun[1], Zijun Zhang[1], Ming Chen[1,*]

[1]College of Life Sciences, Zhejiang University, Hangzhou 310058, China

Summary

Understanding how metabolic reactions translate the genome of an organism into its phenotype is a grand challenge in biology. Genome-wide association studies (GWAS) statistically connect genotypes to phenotypes, without any recourse to known molecular interactions, whereas a molecular mechanistic description ties gene function to phenotype through gene regulatory networks (GRNs), protein-protein interactions (PPIs) and molecular pathways. Integration of different regulatory information levels of an organism is expected to provide a good way for mapping genotypes to phenotypes. However, the lack of curated metabolic model of rice is blocking the exploration of genome-scale multi-level network reconstruction. Here, we have merged GRNs, PPIs and genome-scale metabolic networks (GSMNs) approaches into a single framework for rice via omics' regulatory information reconstruction and integration. Firstly, we reconstructed a genome-scale metabolic model, containing 4,462 function genes, 2,986 metabolites involved in 3,316 reactions, and compartmentalized into ten subcellular locations. Furthermore, 90,358 pairs of protein-protein interactions, 662,936 pairs of gene regulations and 1,763 microRNA-target interactions were integrated into the metabolic model. Eventually, a database was developped for systematically storing and retrieving the genome-scale multi-level network of rice. This provides a reference for understanding genotype-phenotype relationship of rice, and for analysis of its molecular regulatory network.

1 Introduction

Rice, one of the world's most important crops, has important syntonic relationships with other cereal species and is a model plant for grasses study [1]. Its compact genome size (≈430 Mb), well-established methods for genetic transformation, availability of high-density genome maps both genetically and physically [2-4], and finished genome annotation, all make rice an ideal model systems to study plant physiology, development, agronomics and genomics of grasses [5,6].

In recent years, several laboratories have successfully developed different omic' level models of rice, including gene regulatory networks, small RNA mediated gene regulatory networks, protein interaction networks, etc. Lee *et al.*'s study proposed an experimentally tested genome-scale gene network of rice based on 24 data types [7]. In our previous work, a very reliable microRNA-mediated gene regulatory network of rice was predicted and further filtered based on degradome sequencing data [8]. We also developed the first well annotated protein protein interactome network of rice, PRIN [9]. It has greatly extended the current available protein-protein interaction data of rice, which will certainly provide further insights into rice functional genomics and systems biology.

* To whom correspondence should be addressed. Email: mchen@zju.edu.cn

Copyright 2013 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

In 1995, the first genome (*Haemophilus influenzae* Rd) was completely sequenced [10], and in 1999, the first genome scale metabolic model was reconstructed by Jeremy S. Edwards *et al.* [11]. Up to date, more than 135 genome scale metabolic models were reconstructed, involving about 90 species. Among them, there are only two models for plant species (*Arabidopsis, Zea mays*) [12-14]. The model by Dal'Molin *et al.* [12] has identified the set of essential reactions, accounting for the classical photorespiratory cycle, and has highlighted the significant differences between photosynthetic and non-photosynthetic metabolites. The model of Poolman *et al.* [13] includes ATP demand constraints for biomass production and maintenance, suggesting strategies for the construction of metabolic modules as a consequence of variation in ATP requirement. The first attempt of globally characterizing the metabolic capabilities (both primary and secondary metabolism) with a compartmentalized photosynthetic model was implemented on *Zea mays* iRS1563, an important crop and energy plant [14], while genome-scale metabolic reconstruction of rice is still to be explored.

Recently, several studies have proposed strategies for the genotype-phenotype mapping. Sergey V. Nuzhdin *et al.* [15] suggested that Genome wide association studies (GWAS) and Gene Regulatory Networks (GRNs) should be merged together to elucidate and quantify the molecular pathways underlying phenotypic variation. It is clear that approaches of integrating genetics and omics would be a valuable strategy for investigating the regulation of the relationship between plant metabolism and physiology. In previous research, GWAS were able to statistically connect genotypes to phenotypes, without any recourse to known molecular interactions, whereas a molecular mechanistic description ties gene function to phenotype, through GRNs, protein-protein interactions (PPIs) and molecular pathways. The potential of metabolomics as a functional genomics tool in addition to transcriptomics and proteomics is well recognized [16]. Therefore, the integration of different levels of regulatory information (genome, proteome and metabolome) could probably be a new approach for mapping genotypes to phenotypes. The determination of rice genome sequence (and its annotation), of proteome interactions and of transcriptome regulatory information, have led to the accumulation of sufficient public data to construct systems-level models. These models could increase the understanding of genotype-phenotype relationship, and consequently help to improve the quality and productivity of rice. However, the lack of curated metabolic reconstruction of rice prevents the construction of genome-scale multi-level network. Here, we present, for the first time, a reconstructed and curated genome-scale metabolic model of rice, including gene regulatory network, microRNA-targets information and protein-protein interactions (Figure 1). The genome-scale multi-level networkprovides a detailed reference for rice molecular regulatory analysis and genotype-phenotype mapping. Eventually, a comprehensive molecular regulation database of rice has been developing to systematically store, analyze and visualize the rice genome-scale multi-level network.

2 Database construction

2.1 2.1 Data Source

The annotated genome of rice was retrieved from the Rice Genome Annotation Project [17]. Kyoto Encyclopedia of Genes and Genomes (KEGG) [18], RiceCyc [19], UniProt [20] and Brenda [21] provided gene-enzyme-reaction associated information. Metabolites identifiers were extracted from PubChem Compounds [22] and Chemical Entities of Biological Interest (ChEBI) databases [23]. The gene network [7] and the microRNA-mediated gene regulatory network of rice [8] provided insights into the related gene regulatory network. Protein-protein interactions network was established using data from PRIN [9], BIND [24] and PlaPID [25]. All data sources mentioned above are summarized in Table 1.

Copyright 2013 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

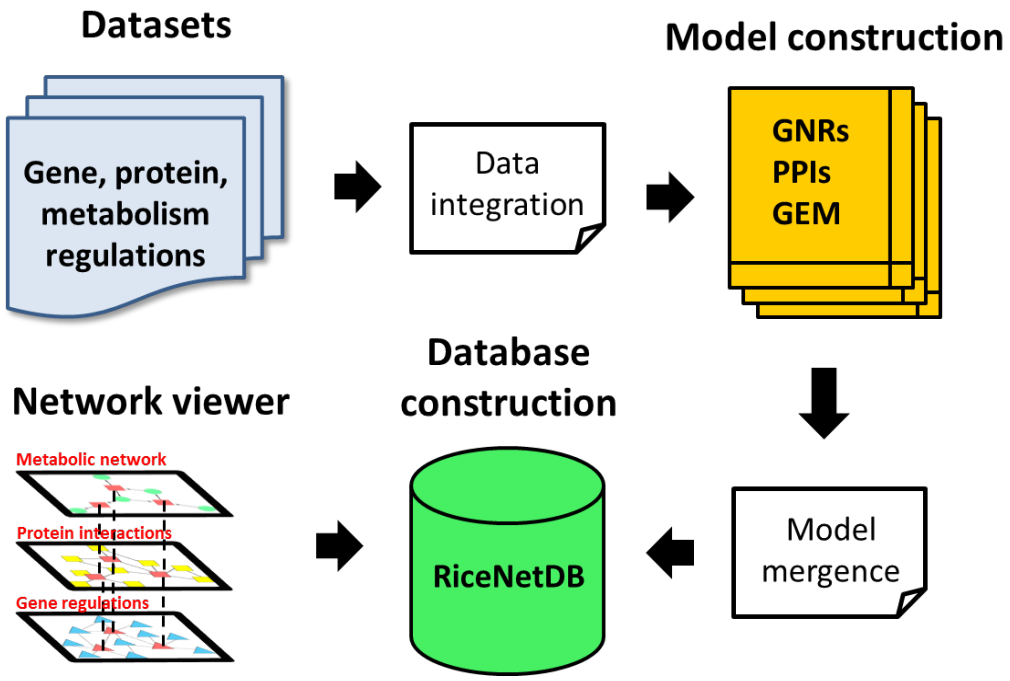

Figure 1: A workflow of genome-scale multi-level network reconstruction of rice.

2.2 Metabolic reconstruction

A genome-scale metabolic model of rice (RiceGEM) was constructed covering primary metabolism for a compartmentalized plant cell based on the rice genome. RiceGEM is a curated genome-wide metabolic model with 4,462 functional genes, 2,986 metabolites and 3,316 reactions, compartmentalized into ten subcellular locations.

2.2.1 Draft reconstruction

A gene-centric organization of metabolic information was adopted, in which each known metabolic gene was mapped to one or several reactions. The reconstruction process was semi-automatic. The procedure previously applied to the genome-scale metabolic reconstruction of Arabidopsis [26] was used to integrate the gene-enzyme-reaction associated information. The first step was to create a list of common metabolites by merging together KEGG, PubChem, ChEBI, and SMILES identifiers. The second step was to find the reactions associated with each enzyme/metabolite pair and to identify new compounds (matched compounds) in these reactions. The whole process was repeated several times until no additional matched compounds could be found. The reconstructed draft metabolic network contains 4,462 metabolic functional genes, a total of 3,316 unique reactions (including 1,372 from KEGG, 1,926 from RiceCyc, 411 from Uniprot and 242 from Brenda) and 2,986 metabolites (including 1,484 from KEGG, 1863 from RiceCyc, 848 from Uniprot and 481 from Brenda).

2.2.2 Manual Curation

Most reactions retrieved from the public database were not balanced. Protons and water were frequently omitted from reaction stoichiometry. And charge protonation state of the compounds did not correspond with intracellular pH. In our study, Marvin (version 5.3.3, ChemAxon Kft) was used to calculate the net charge of individual metabolites at pH 7.2. This pH was assumed to be the same for all organelles. Then, against the linear algebra, all reactions were balanced based on the new charges of the metabolites.

Copyright 2013 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Table 1: Online resources for the reconstruction of rice multi-level network

Database	Link
Genome database	
Rice Genome Annotation Project (TIGR)	http://rice.plantbiology.msu.edu/
Pathway database	
Kyoto Encyclopedia of Genes and Genomes (KEGG)	http://www.genome.jp/kegg/pathway.html
RiceCyc Version 3.3	http://pathway.gramene.org/gramene/ricecyc.shtml
Uniprot	http://www.uniprot.org/
Brenda	http://www.brenda-enzymes.info/
Enzymes databases	
ExPASy Enzyme Database	http://www.expasy.org
Brenda	http://www.brenda-enzymes.info/
Compounds database	
PubChem Compounds	http://pubchem.ncbi.nlm.nih.gov/
Chemical Entities of Biological Interest (ChEBI)	http://www.ebi.ac.uk/chebi/
Gene regulatory database	
Probabilistic Functional Gene Network of *Oryza sativa*	http://www.functionalnet.org/ricenet/
Protein-protein interaction database	
a Predicted Rice Interactome Network (PRIN)	http://bis.zju.edu.cn/prin/
the Biomolecular Interaction Network databank (BIND)	http://bind.ca
A Database of Protein-Protein Interactions in Plants (PlaPID)	http://www.plapid.net/

Due to the fact that today there is relatively little thermodynamics data in reaction databases, the free energies of the reactions in physiologic conditions (298.15 K, pH 7.2, and 1mM concentrations of all species, except for H^+ and water) were computed in two steps. First, all free energies of reactions at standard conditions (1 atm, pH 7, 298.15 K, zero ionic strength and 1M concentrations of all species except H^+ and water) were estimated ($\Delta_r G_{est}^{'0}$) by the group contribution method. Then the $\Delta_r G_{est}^{'0}$ were adjusted to physiologic conditions by the method [27].

At present, exsiting relevant databases contain little information on the proteome subcellular localization of rice. In the absence of sound experimental information, we predicted the proteins localization from the amino acid sequences, using the following publicly available softwares: CELLO [28], epiloc [29], mPloc [30], Predotar [31], TargetP [32], Wolf PSORT [33], subcellPredict [34] and PROlocalizer [35]. Two or more consistent localization results were accepted as a consensus prediction. As a result, the rice genome scale metabolic model was compartmentalized into ten subcellular locations. Specifically, mitochondrion includes 8,129 proteins; vacuole 335; golgi apparatus 1,960; cytoplasm 34,088; Endoplasmic Reticulum 4,703; extracellular 3,226; nucleus 7,507; chloroplast 2,800; plasma membrane 640 and peroxysome 2,847.

Copyright 2013 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

2.3 Multi-level network construction

2.3.1 GRNs model

The model of GRNs must include regulations by coding and non-coding genes. Here, the experimentally tested genome-scale gene network of rice proposed by Lee *et al*. [7], was used to build our the network of coding gene regulations. Further, microRNA-target regulations were drawn from the microRNA mediated gene regulatory network of rice, based on degradome sequencing data, from Meng *et al.* [8]. Consequently, the coding gene regulatory network and the microRNA-target information were merged via TIGR ID, resulting in the construction of rice GRNs model.

2.3.2 PPIs model

BIND and PlaPID are repositories of experimentally verified protein-protein interaction information. PRIN is a predicted rice proteome interaction network that greatly extended the knowledge of protein-protein interaction. Our PPIs model was constructed by merging these three datasets, on the base of the TIGR ID.

2.3.3 Model integration

The genome-scale multi-level network of rice was created by the integration of RiceGEM, GRNs and PPIs. Since the integrated data include different types of identifiers, cross-identification tables were constructed by converting all interactor IDs into UniProtKB AC, TIGR ID and compounds ID (developed by us). Then, the interaction tables describing gene-gene, gene-protein, enzyme-compound, enzyme-reaction, compound-reaction and reaction-pathway relations were built by ID mapping.

3 Database implementation

Our rice multi-level network database has been implemented in MySQL, and can be queried by a web interface (Figure 2). In current version, it stores 4,462 metabolic associated genes, 2,986 metabolites, 3,316 reactions, 90,358 pairs of protein-protein interactions, 662,936 pairs of gene regulations and 1,763 microRNA-target interactions. The database covers frequently-used information about genes (names, attributions, ontologies, genes regulators descriptions), proteins (interactors, subcellular locations) and compounds (names, reactions, pathways, evidences). Much effort was put on providing a convenient data access for both occasional and advanced users, *i.e.* all IDs for cross-reference were preserved.

The database can be searched by multiple types of gene/protein/compound/reaction/pathway IDs, including UniProtKB AC, TIGR ID, Gene Symbol, KEGG compounds/reaction/pathway ID and RiceCyc compounds/reaction/pathway ID. A sequence search service (BLAST search) is also available. The result pages contain not only the interaction information, but also the annotation and classification for the query entries as well as the interacting partners. In addition, we have developed four browser modules (Genome Browser, Browses by Enzyme, Browses by Pathway and Browses by Subcellular location) for Genome, Enzymes, Pathways and Subcellular locations. Furthermore, a viewer is available for presenting an overview of the multi-level network (gene/protein/metabolic networks) or a fragment of it, associate to a queried gene or protein. These two views are presented in Figure 3.

Copyright 2013 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

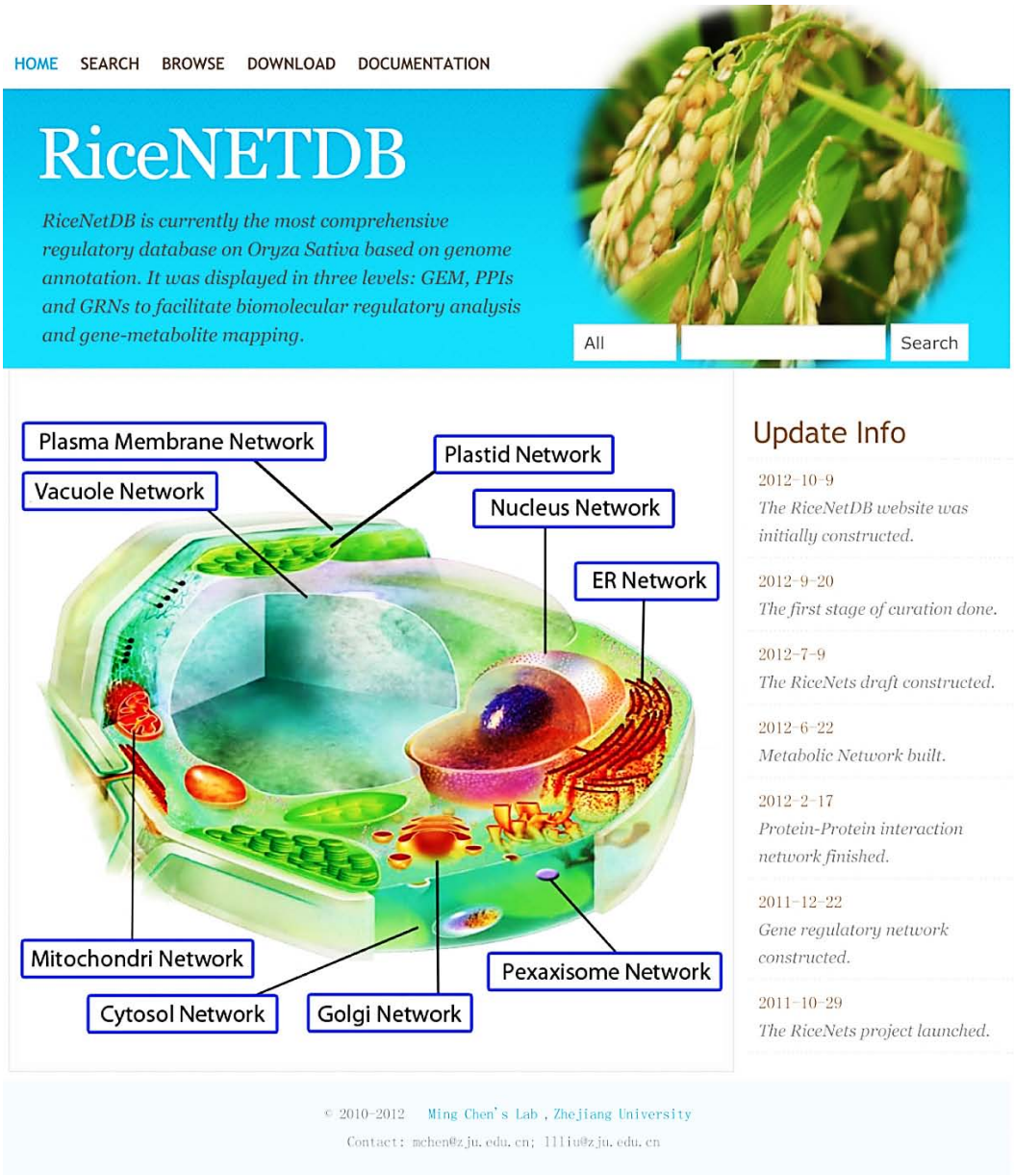

Figure 2: The RiceNetDB website screenshot. A public accessible version of RiceNetDB is scheduled to release later 2013.

The database is downloadable under a certain agreement for further study. Submission of new experimentally tested data is highly welcomed and appreciated, through a special web interface. After validation, the data will be integrated into our multi-level network.

4 Conclusion

We have presented an integrative and specialized database for biomolecular regulatory analysis and gene-metabolite mapping. It will be timely updated with newly available interactions information. In the future, we also intend to integrate expression data and tissue specific information into this model. With the development of rice omics' regulatory studies, our database could be enriched with these new experimental data. In the study of genotype-phenotype relationship, more and more attention is paid to the dynamic path from genotypes to phenotypes through gene regulatory networks, protein interactions and metabolic networks. In the future, the results of these new researches will be integrated into our database.

Copyright 2013 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

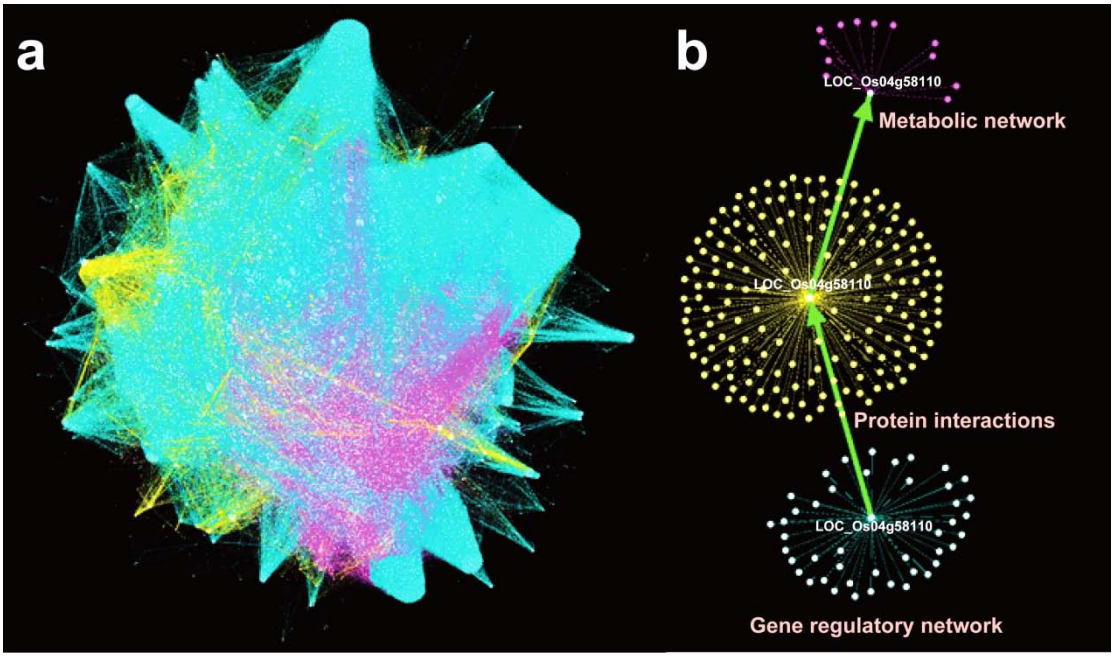

Figure 3: The genome-scale multi-level network of rice; (a) overview of the genome-scale multiple level network of rice visualized by Gephi (http://www.gephi.org), blue color denotes the gene regulations, yellow denotes the protein-protein interactions and pink shows the protein-metabolite interactions; (b) view of gene centric (LOC_Os04g58110) multi-level network.

Acknowledgements

The authors would like to thank all the publicly available data sets and the scientists behind them. This work was supported by the National Natural Sciences Foundation of China [30771326, 30971743], the Program for New Century Excellent Talents in University of China [NCET-07-0740].

References

[1] T. Sasaki and B. Burr. International rice genome sequencing project: the effort to completely sequence the rice genome. Current Opinion in Plant Biology, 3:138-141, 2000.

[2] Y. Harushima, et al. A high-density rice genetic linkage map with 2275 markers using a single F2 population. Genetics, 148:479-494, 1998.

[3] J. Wu, et al. A comprehensive rice transcript map containing 6591 expressed sequence tag sites. Plant Cell, 14:525-535, 2002.

[4] M. Chen, et al. An integrated physical and genetic map of the rice genome. Plant Cell, 14:537-545, 2002.

[5] M. D. Gale and K. M. Devos. Comparative genetics in the grasses. Proceedings of the National Academy of Sciences, 95:1971-1974, 1998.

[6] S. A. Goff. Rice as a model for cereal genomics. Current Opinion in Plant Biology, 2:86-89, 1999.

Copyright 2013 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

[7] I. Lee, et al. Genetic dissection of the biotic stress response using a genome-scale gene network for rice. Proceedings of the National Academy of Sciences, 45:18548-18553, 2011.

[8] Y. J. Meng, et al. Construction of microRNA- and microRNA*-mediated regulatory networks in plants. RNA Biology, 8(6):1124-1148, 2011.

[9] H. B. Gu, et al. PRIN, a predicted rice interactome network. BMC Bioinformatics, 12:161, 2011.

[10] R. D. Fleischmann, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science, 269(5223):496-512, 1995.

[11] J. S. Edwards and B. O. Palsson. Systems Properties of the Haemophilus influenzae Rd Metabolic Genotype. Journal of Biology Chemistry, 274:17410-17416, 1999.

[12] M. G. Poolman, et al. A Genome-Scale Metabolic Model of Arabidopsis and Some of Its Properties. Plant Physiology, 151:1570-1581, 2009.

[13] C. G. de O. Dal'Molin, et al. AraGEM, a Genome-Scale Reconstruction of the Primary Metabolic Network in Arabidopsis1. Plant Physiology, 152:579-589, 2010.

[14] R. Saha, et al. Zea mays i RS1563: A Comprehensive Genome-Scale Metabolic Reconstruction of Maize Metabolism. PLoS ONE, 6:e21784, 2011.

[15] S. V. Nuzhdin, et al. Genotype–phenotype mapping in a post-GWAS world. Trends in Genetics, 28(9):421-426, 2012.

[16] N. Carreno-Quintero, et al. Genetic analysis of metabolome–phenotype interactions: from model to crop species. Trends in Genetics, 29:41-50, 2012.

[17] S. Ouyang, et al. The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Research, 35:D883-D887, 2007.

[18] M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research, 27(1):29-34, 1999.

[19] C. Z. Liang, et al. Gramene: a growing plant comparative genomics resource. Nucleic Acids Research, 36:D947-D953, 2008.

[20] R. Apweiler, et al. UniProt: the Universal Protein knowledgebase. Nucleic Acids Research, 32:D115-D119, 2004.

[21] I. Schomburg, et al. BRENDA, enzyme data and metabolic information. Nucleic Acids Research, 30(1):47-49, 2002.

[22] Y. L. Wang, et al. PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Research, 37:W623-W633, 2009.

[23] K. Degtyarenko, et al. ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Research, 36:D344-D350, 2008.

[24] G. D. Bader, et al. BIND: The Biomolecular Interaction Network Database. Nucleic Acids Research, 29(1):242-245, 2001.

[25] S. Mao, et al. Plant Protein Interaction Database (PlaPID): A Comprehensive Analysis and Genome-wide Prediction for Arabidopsis Protein Interactome. (Under Review)

[26] K. Radrich, et al. Integration of metabolic databases for the reconstruction of genome-scale metabolic networks. BMC Systems Biology, 4:114, 2010.

[27] R. A. Alberty. Thermodynamics of Biochemical Reactions. Massachusetts Institute of Technology, MA , 2003.

Copyright 2013 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

[28] C. S. Yu, et al. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Science, 13:1402-1406, 2004.

[29] S. Brady, and H. Shatkay. EpiLoc: A (Working) Text-Based System for Predicting Protein Subcellular Location. Pacific Symposium on Biocomputing, pp. 604-615, 2008.

[30] K. C. Chou and H. B. Shen. Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization. PLoS ONE, 5:e11335, 2010.

[31] I. Small, et al. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics 4(6):1581-1590, 2004.

[32] O. Emanuelsson, et al. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. Journal of Molecular Biology, 300:1005-1016, 2000.

[33] P. Horton, et al. WoLF PSORT: protein localization predictor. Nucleic Acids Research, 35:W585–W587, 2007.

[34] B. Niu, et al. Using AdaBoost for the Prediction of Subcellular Location of Prokaryotic and Eukaryotic Proteins. Molecular Diversity, 12:41-45, 2008.

[35] K. Laurila and M. Vihinen, PROlocalizer: integrated web service for protein subcellular localization prediction. Amino Acids, 40(3):975-80, 2011.

Copyright 2013 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).