MADMAX – Management and analysis database for multiple ~omics experiments

Ke Lin[1,2], Harrie Kools[3], Philip J. de Groot[4], Anand K. Gavai[1,5], Ram K. Basnet[2], Feng Cheng[6], Jian Wu[6], Xiaowu Wang[6], Arjen Lommen[3], Guido J. E. J. Hooiveld[4], Guusje Bonnema[2], Richard G. F. Visser[2], Michael R. Muller[4], Jack A. M. Leunissen[1,5*]

[1] Laboratory of Bioinformatics, P.O. Box 569, 6700 AN Wageningen, the Netherlands

[2] Laboratory of Plant Breeding, P.O. Box 386, 6700 AJ Wageningen, the Netherlands

[3] RIKILT, P.O. Box 230, 6700 AE Wageningen, the Netherlands

[4] Nutrition, Metabolism and Genomics Group, P.O. Box 8129, 6700 AA Wageningen, the Netherlands

[5] Netherlands Bioinformatics Centre (NBIC), P.O. Box 9101, 6500 HB Nijmegen, the Netherlands

[6] Institute of Vegetable and Flowers, Chinese Academy of Agricultural Sciences, 12 Zhongguancun South Street, Beijing, 100081, China

Summary

The rapid increase of ~omics datasets generated by microarray, mass spectrometry and next generation sequencing technologies requires an integrated platform that can combine results from different ~omics datasets to provide novel insights in the understanding of biological systems. MADMAX is designed to provide a solution for storage and analysis of complex ~omics datasets. In addition, analysis results (such as lists of genes) can be merged to reveal candidate genes supported by all datasets. The system constitutes an ISA-Tab compliant LIMS part, which is linked to the different analysis pipelines. A pilot study of different type of ~omics data in *Brassica rapa* demonstrates the possible use of MADMAX. The web-based user interface provides easy access to data and analysis tools on top of the database.

1 Introduction

To better understand how phenotypes emerge, increasingly series of ~omics technologies (genomics, transcriptomics, proteomics, metabolomics) rather than individual measurements are necessarily used within a single study. Such efforts boost the demands of both data storage and data analysis of different high-throughput approaches. However, in the past it was hardly possible to store metadata from different ~omics technologies in the same repository. To accommodate this demand the ISA-Tab [1] format was proposed to build up a common structured representation of the metadata of studies from a combination of technologies. This also triggered attempts to develop data processing tools tailored to the needs of biologists. Unfortunately most of these tools have high demands on hardware requirements, or contain non-intuitive command line-based interfaces.

* To whom correspondence should be addressed: jack.leunissen@wur.nl

doi:10.2390/biecoll-jib-2011-160

Here we present MADMAX, a multi-purpose database for the management and analysis of data from multiple ~omics experiments. It includes an ISA-Tab compliant backend database and a series of analysis pipelines for transcriptomics, metabolomics and genomics datasets; these pipelines are connected to the database through webservices such that other pipelines can be easily integrated into the current system (Figure 1). The currently supported pipelines include gene function annotation for next generation sequencing of genomes, quality control and statistical analysis (such as Gene Set Enrichment Analysis [2], GSEA) using R [3] and Bioconductor [4] for different microarray platforms, and the Metalign software [5] for LCMS, GCMS and GC-GC-MS metabolomics studies. Because the quality of the gene function annotation is the key to reliable transcriptomics and metabolomics analysis in newly sequenced species, MADMAX uses the ProGMap database for orthologous group information [6] to obtain the function annotation for each gene. Besides the original output from microarray and metabolomics analyses, it will further mine the results and generate over- and under-expressed genes from microarray studies and genes responsible for producing enzymes that affect steps in the pathways of metabolites detected in metabolomics studies. The intersection of genes listed in different omics results may lead to a manageable number of candidate genes for experimental validation.

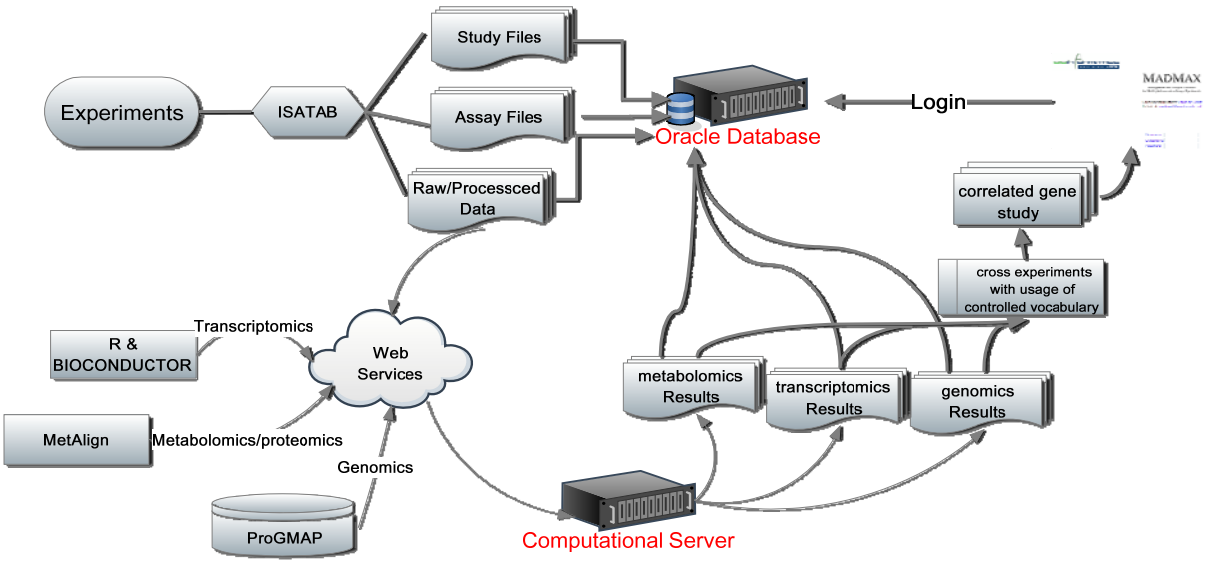

Figure 1: Architecture and pipeline of MADMAX

Through the web interface, the user can store a complete experiment with all fields required in ISA-Tab format, sufficient to allow for subsequent analysis or even repeating the experiment later. Another section on the website is the central access to different analysis pipelines. Both individual analysis results and combined gene lists can be retrieved in the system for download. Centrally stored experiments and analysis results can only be accessed by the creator by default and will be accessible for other users only if the creator desires to share the data. The system is on an automatic backup schedule.

MADMAX can be reached at http://madmax2.bioinformatics.nl/ and is available upon request by sending an email to madmax.request@bioinformatics.nl.

2 Implementation

MADMAX is built upon an Oracle relational database on a Linux server and a computational analysis engine for different ~omics data on a second server. The transcriptomics and

metabolomics analysis pipelines are triggered by web services through a web-based user interface developed using Oracle Application Express. The genomics analysis is done separately outside of the database system, and loaded to the system when it's done. Both metadata and analysis results of an experiment are stored in the database in ISA-Tab compliant format. Different ~omics analysis results can be further combined to yield system-level measurements within one experiment.

2.1 Database infrastructure

The basic unit in MADMAX is the Experiment which contains the overall goals and means used in an experiment as with Investigation in the data model of the ISA-Tab standard. Each experiment can have one or more studies to record the sample preparation information. One study can be used for one or more assays if these assays share the same sample preparation. In an assay, each sample and the corresponding data file generated from it will be connected with additional information about the extracted material; the processed results from the computational server in MADMAX will be attached to the assay where the raw data are used in the analysis.

2.2 Analysis pipeline

Currently MADMAX can handle three types of ~omics datasets: gene models of a partial or complete genome, microarrays from Affymetrix, Illumina, or Agilent, and LC/GC/GCGC mass spectrometry data sets. By default, a list of genes will be generated from the different ~omics data in an experiment, based upon gene models from genomics, over/under expressed genes from microarrays, and genes encoding the enzymes involved in the pathways of the metabolites detected in metabolomics analysis. Not all analysis pipelines used in the system can be accessed by the user interactively. As for genomics analysis, the complete gene models will be loaded in the system and annotated later by the system maintainer. Both microarray and metabolomics analyses can be accessed interactively through the web interface.

2.2.1 Genomics analysis

The *de novo* sequenced genomes using next-generation technology normally contain *in silico* predicted gene models. The gene models in FASTA format are then used as query sequences to BLAST against the ProGMap protein database, which includes protein sequences from Ensembl version 61_8, InParanoid version 7, OrthoMCL version 4, COG/KOG from 2003, eggNOG version 2, HomoloGene version 65, ProtClustDB from Nov 2010, PIRSF from Mar. 9 2010, OMA from Nov 2010 and KEGG ORTHOLOGY database [7-16] from Apr 2011 (at the time of this writing). At present, 21,521,041 proteins sequences are used in the search with default BLASTX [17] settings. When the BLAST search finishes, the matched sequences are filtered through three thresholds: matched sequence length is greater than 100 amino acids, percentage of sequence similarity is greater than 40% and the e-value of the sequence is lower than 10E-20; the results are reformatted into a tab delimited file. Gene models will then be mapped to an orthologous group of databases mentioned above one by one using passed sequences. The criteria of orthologous group selection are mainly based on four features: database source, group evidence level, number of matched sequences in the orthologous group and sequence evidence level. Specifically, orthologous groups from the database with highest priority score in Supplementary Table S1 will be taken into account in the first place. Priority scores assigned to each database are determined on the basis of their curator, the number of protein sequences used, and the frequency of database release. The Ensembl and InParanoid databases get relatively low priority due to their pairwise comparison implementation, which may reduce the reliability for some conserved function group