Project No. *222886-2*

MICROME

The Microme Project:
A Knowledge-Based Bioinformatics Framework for Microbial Pathway Genomics

Instrument: **Collaborative project**

Thematic Priority: **KBBE-2007-3-2-08: BIO-INFORMATICS - Microbial genomics and bio-informatics**

D2.3 Microme data set release 1.0 (first public release)

Due date of deliverable: 1 Oct 2011
Actual submission date: 25 Oct 2011

Start date of project: 1.12.2009 Duration: 48 months

Organisation name of lead contractor for this deliverable: CEA

Project co-funded by the European Commission within the Seventh Framework Programme (2009-2013)		
	Dissemination Level	
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Contributors

The Microme release data set has been produced primarily by the CEA, EBI, and SIB, with additional contributions from the other Microme partners.

INTRODUCTION

D2.3 Microme data set release 1.0 (first public release)

The aim of this deliverable was the production of a complete data set for Microme, suitable for inclusion in the first public release. This involves the curation of the pre-release data set, the production of pathway diagrams, and the implementation of the pathway projection algorithm previously developed, and the testing and revision of pathways as appropriate through the use of the Microme web portal (which was in development, and available to project partners, during the period in which the data set was prepared).

Methods

The production of the release dataset (following from the previous reporting of D2.2) required the following work:
(i) Familiarisation of project members with Microme curation tools. This was initiated at the first Microme annotation jamboree (D7.1), and continued thereafter by remote interaction (using screensharing tools, e.g. Skype, where appropriate) between project members.
(ii) Production of pathway diagrams for all pathways to be incorporated in the first production release.
(iii) Analysis of imported pathways using the pathway diagrams; this led to
A. The correction/improvement of the initial data import pipelines
B. The identification of irregularities in the imported data; the determination of policies to mediate these irregularities (to be incorporated in the evolving Microme Standard Operating Procedures (D2.4); and the interaction with external data suppliers as appropriate.
C. The fixing of specific pathways in accordance with the outcomes of B, above
(iv) The initial production run of the Microme pathway projection process, involving
A. The production of genomic databases for the initial 30 species to which pathway projection has been applied.
B. The determination of inferred orthology relationships from the genomic data.
C. The inference of catalytic functions, and of the presence/partial presence of pathways in the recipient species
D. The integration of the propagated data into the Microme database.
(v) The establishment of a procedure for the ongoing curation, and periodic release, of Microme data.

Results (if applicable, interactions with other workpackages)

The dataset has been curated with the aid of the Microme curation interface developed as deliverable D1.3. This interface allows for the production of pathway diagrams, and for the editing of pathways (the insertion/modification of products, substrates, catalysts, cross-references etc.). However, certain bulk changes have been applied to many pathways through the direct modification of the database.

The examination of the pathways imported into the pre-release data set (D2.2) took place in parallel with the production of pathway diagrams for each pathway. The operation of the Reactome software is centred on a visual display of pathways, drawn using SBGN (Systems Biology Graphical Notation), a standard developed for the schematic representation of biological systems. Reactome editing software, also adapted for use by Microme, supports the drawing of such diagrams. Layout is potentially complex, especially where a diagram is indicating the existence of a transportation reaction (involving a change in molecular localisation). The software attempts an automatic layout of the pathway, which must then be edited by a curator into a usable form. The process of editing pathway diagrams provides a rigorous quality control of the data in the pathway, given that inconsistencies in the data lead to obvious visual anomalies in the pathway. Such anomalies were identified, grouped, diagnosed and manual and/or automatic fixes proposed. In many cases, the fixes required included a manual fix of the data being prepared for release, together with a change to the import procedures to correct errors in the prototype software or to perform more extensive checks for errors or inconsistencies in the data being imported. The process was managed using the enterprise wiki Confluence, the JIRA bug tracking system, and multi-way phonecalls between the partners. Confluence was used to assign intiail work, JIRA was used to document problem cases, confluence was used for further characterisation of the problems, general issues were then discussed by phone and by mail, then as specific issues were solved, the appropriate JIRA tickets were closed.

In summary, particular issues included:

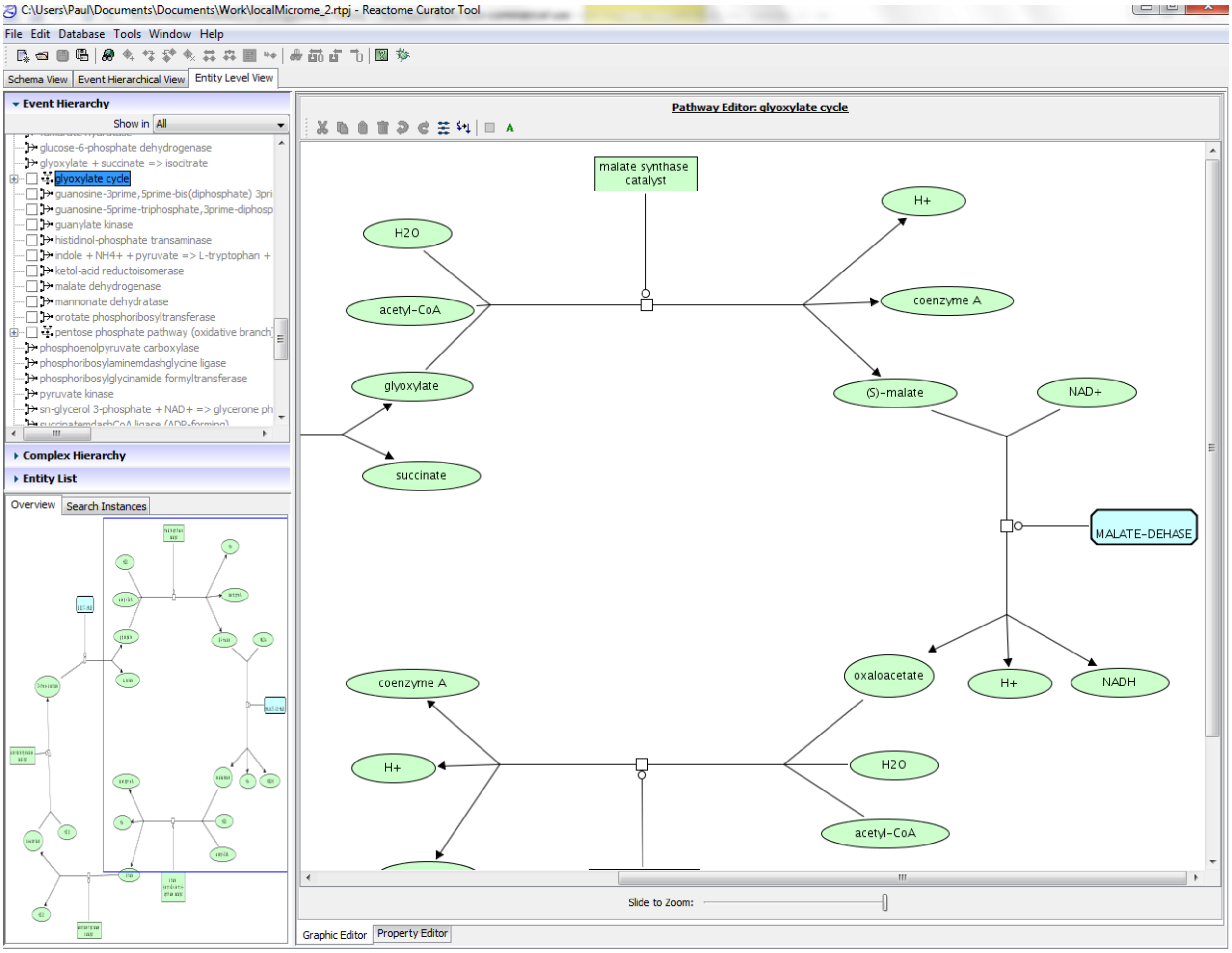
(i) pathways that switched from generic to specific forms, and vice versa. This resulted, in some cases, from insufficiently precise resolution of different primary data sources in the initial data import process (i.e. some cross-referenced entities may describe more/less specific forms of reactions, and participating moleciles, than other data sources); in some cases, from inconsistency within individual primary sources; and in some cases, from differences in the data model used in Microme and that used in the primary sources.

(ii) multiple database entities representing the same real world entity. This was caused by inconsistent localisation or chemical identification of entities to compartments in a source database; errors in localisation or chemical identification introduced during the import process; the existence of alternative ways of representing a generic entity in the Microme data model, with insufficient information sometimes available in the external sources to allow the consistent choice of the appropriate form in all cases; or the failure to represent all chemicals in their most common form at ph7.3.

(iii) Issues to do with names displayed, which required correction and modification to appropriate parsers.

Many of these issues required extensive analysis and diagnosis before an appropriate fix could be applied. A presentation that reviewed some of the outstanding issues is attached as the first appendix to this report. In total, 330 reference pathways have been checked, and had pathway diagrams drawn following the resolution of outstanding intra- and inter-pathway issues. Figure 1, below, demonstates the creation of a diagram in the curation interface,

Figure 1 Creation of a pathway diagram for the glyoxylate cylcle in *Escherichia coli.*


C:\Users\Paul\Documents\Documents\Work\localMicrome_2.rtpj - Reactome Curator Tool
File Edit Database Tools Window Help
Schema View
Event Hierarchical View
Entity Level View
Event Hierarchy
Show in All
glucose-6-phosphate dehydrogenase
glyoxylate + succinate => isocitrate
glyoxylate cycle
guanosine-3prime,5prime-bis(diphosphate) 3pri
guanosine-5prime-triphosphate,3prime-diphosp
guanylate kinase
histidinol-phosphate transaminase
indole + NH4+ + pyruvate => L-tryptophan +
ketol-acid reductoisomerase
malate dehydrogenase
mannonate dehydratase
orotate phosphoribosyltransferase
pentose phosphate pathway (oxidative branch)
phosphoenolpyruvate carboxylase
phosphoribosylaminemdashglycine ligase
phosphoribosylglycinamide formyltransferase
pyruvate kinase
sn-glycerol 3-phosphate + NAD+ => glycerone ph
Complex Hierarchy
Entity List
Overview
Search Instances
Pathway Editor: glyoxylate cycle
malate synthase catalyst
H+
H2O
coenzyme A
acetyl-CoA
glyoxylate
(S)-malate
NAD+
succinate
MALATE-DEHASE
oxaloacetate
H+
NADH
coenzyme A
H+
H2O
acetyl-CoA
Slide to Zoom:
Graphic Editor
Property Editor

To enable pathway projection, 30 species were identified for inclusion in the first release, on grounds of scientific interest and taxonomic spread. These 30 species have also been targeted for experimental study using Biolog technology in work package 3, which will allow the remediation of theoretical pathways with actual metabolic flux data in the later stages of the project. 24 of these species were in an appropriate state of annotation for inclusion in the first release. In the first step of work, a genomic database was constructed containing these species using the Ensembl software framework (Flicek, P *et al. Nucleic Acids Res*. **2010** 38: D557-62). Ensembl is a robust, well-proven framework for representing genomic data; the data source itself was the genomic sequence and annotations submitted for each of these genomes into the international public nucleotide sequence archives. The database produced is attached to this deliverable as appendix D2.3.2.

A procedure for pathway projection was derived from that used in the Reactome project. This involves several steps: the first is the identification of orthologous proteins between species; the second is the derivation of reaction catalysts (and hence pathways) in non-reference species; the third is the updating of the database to include the new derived pathways (including the automatic generation of diagrams derived from the manually drawn diagrams of the reference pathways.

To carry out this procedure, the master database was frozen and a copy made, to prepare for release. The Ensembl Compara gene trees algorithm (Vilella AJ, *et al. Genome Res*. **2009** Feb;19:327-35) was applied to the genomic database to derive orthologues, and the derived entities added to the release freeze. Curation activities continue on the master database, and new freezes will be made at intervals and projected data re-derived. The comparative genomics database generated is attached to this deliverable as Appendix D2.3.3; the final release database is attached as D2.3.4. All 3 databases are used to support different components of the Microme web interface (see D.6.1), which can be used to explore the data. Summary tables, describing the orthology relationships and pathways inferred, are attached as Appendix D2.3.5. In total, 18476 orthology relationships were inferred across the 23 recipient species. The numbers of wholly or partially projected pathways varied according to the closeness of the recipient species to *Escherichia coli*, from which the initial reference data was derived: from 298/330 in Salmonella enterica subsp. enterica serovar Typhimurium str. SL1344 (another beta enterobacterium) to 159 in Capnocytophaga ochracea DSM 7271.

With this deliverable, milestone M1 of the project is achieved. This enables the achievement of milestone M2, the first public release of the Microme portal, following the work undertaken in work package 6 to develop a web front end to provide public access to the data.

Appendices:
D2.3.1 Powerpoint presentation used in discussion of outstanding Microme data issues.
D2.3.2 Microme 1.0 genome database, MySQL dump.
D2.3.3 Microme 1.0 comparative genomics database, MySQL dump.
D2.3.4 Microme 1.0 pathway database, MySQL dump.
D2.3.5 Summary of inferred orthology relationships and pathways.

Appendices are available for download at the following URL (Size: 959.8 MB):
ftp://ftp.ebi.ac.uk/pub/databases/integr8/tmp/D2.3_appendices.tar.gz

Perspectives

The preparation of the first public release data set for Microme has helped define and refine the partners' understanding of the data and its flow through the Microme system. It is a necessary precursor of the public release of the Microme portal (D6.1); it has contributed to the development of standard operating procedures (D2.4) that will be applied in the production of subsequent data releases (D2.5, D2.6, D2.7), and will inform revision of the initial data import and pathway projection processes (D1.8, D1.9). Two immediate foci of future work will be an assessment of the orthology projections used; and the improvement of gene annotations used in the process through the incorporation of new/replacement annotations manually curated by CEA.