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**

D2.7: Microme data set release 4.0

Due date of deliverable: 30/11/2013
Actual submission date: 13/1/2014

Start date of project: 1.12.2009 Duration: 48 months

Organisation name of lead contractor for this deliverable: SIB

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

SIB, CEA, EMBL-EBI

EMBL-EBI generated the overall genome-reaction set. CEA and SIB were responsible for the creation of specific SSPAs, and the curation of certain of the source data resources used to build the full release.

INTRODUCTION

D2.7: Microme data set release 4.0

We have produced a 4th major release of the Microme genome-reaction-pathway data, following incremental monthly updates to the previously reported data release 3.0 (D2.6).

Methods

The deliverable has been produced by the continued application of computational inference and curational methods described in previous deliverables, for combining validated genome-reaction catalyst instances and reference pathway information with inferred reaction catalyst instances in other species. A computational pipeline has been developed to automatically gather data from the primary sources and infer additional instances in a lazy manner (i.e. only re-computing inferences where new data exists that might influence the result of the calculation (see D1.9 for details). This pipeline has been deployed in a production environment and used to generate regular data updates. The final data set release combines all data generated by the combination of these methods as of November 2013

Results (if applicable, interactions with other workpackages)

We have generated monthly incremental releases of Microme data since reporting on data release 3 in D2.6. The production pipeline was substantially revised with the release of 3.8 to include additional "gap filling" inferences using the pipeline developed in work package 1 (see D1.9). The following tables and charts show the increase in data in Microme due to (i) the increased number of reference genomes available for annotation in the public archives (ii) the increases quantity of data available in the primary data sources (iii) the use of the gap filling pipeline to supplement the primary data imports.

Table 1
Comparison of the Microme Data releases 3.0 and 4.0.

	3.0 Release Nov 2012	***4.0 Release Nov 2013***	***4.0 Release + Gap Filling***
Genomes	4,831	9,544	9,544
Gene-reaction associations	4,206,505	8,133,739	8,540,484
Genome-reaction associations	1,929,068	3,657,984	3,918,595
Avg. reactions/genome	399.31	389.27	410.58

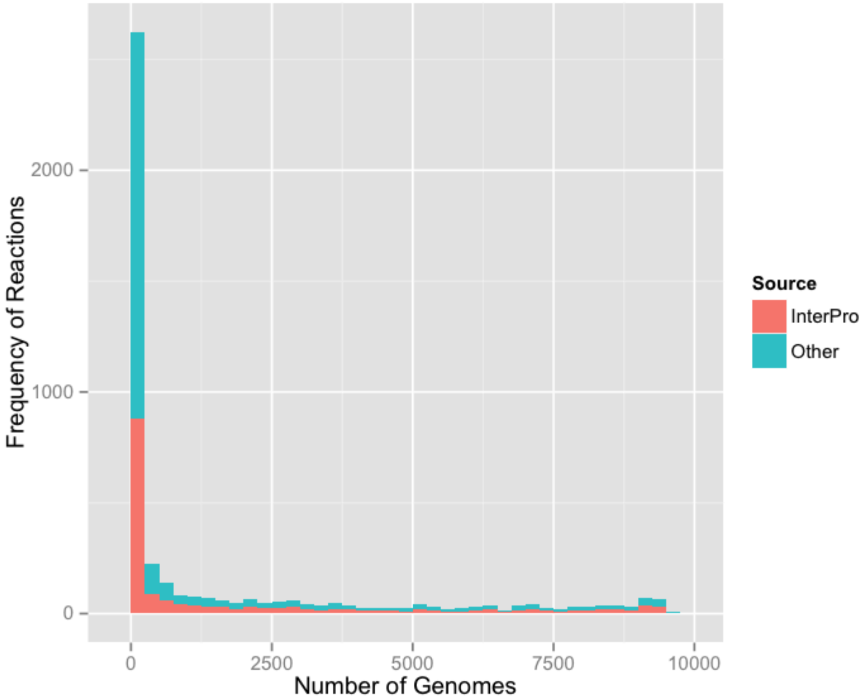

Figure 1. Coverage of the reaction space by genome

The histogram shows the frequency of the number of reactions per genome, with the red bars referring to the reaction-genome instances inferred using on InterPro annotations (the method which identifies the largest number of instances overall). The large number of reactions identified in only a few genomes correspond (to a large extent) to data extracted from the MicroScope platform, in which a smaller number of genomes have been exceptionally well-annotated (including the annotation of reactions for which there is not yet an associated InterPro domain, and so which cannot be inferred more widely). Most genomes are annotated only through automatic methods, which are only capable of identifying a smaller subset of reactions.

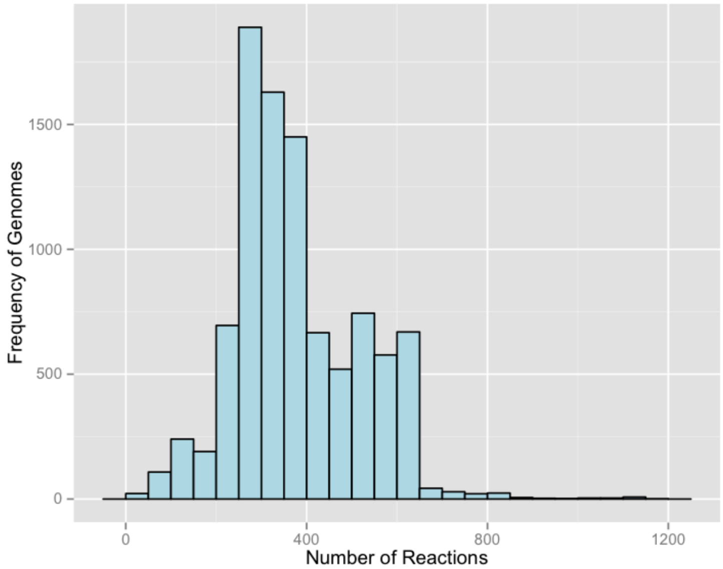

Figure 2. Number of reactions per genome.

The histogram shows the frequency of the number of reactions annotated per genome. Most genomes have been annotated with a approximately 250-600 reactions.

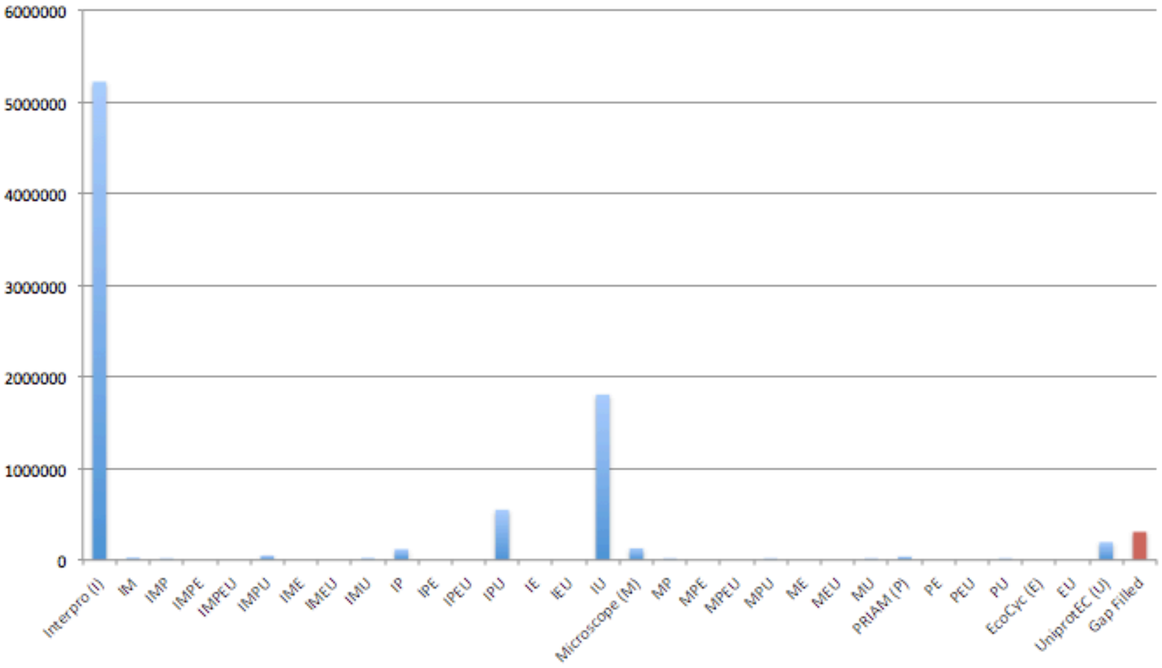

Figure 3. Histogram of the number of unique gene-reaction associations predicted per (combination of) method(s).

Each method is represented with the initial letter. Each bar represents the number of gene- reaction associations predicted together by the combination of methods identified by the corresponding letters.

Key: InterPro (I); Microscpe (M); Priam (P); EcoCyc (combined import and orthology inference) (E); direct UniProt annotation (U). Thus, IMPEU indicates reaction instances identified by all 5 approaches; IM, reaction instances defined by the InterPro-based methods and the MicroScope import only; and o on. Reactions inferred by the gap filling pipeline are shown in red on the right-hand side of the chart.

Species specific pathway assemblies

30 taxonomically diverse genomes were identified at the start of the project for priority attention for curation curation and model building. For these genomes, the goal (in WP2) was to generate a species-specific pathway assembly: a set of all pathways and variants recorded in a genome through a combination of automatic methods and manual curation. These SSPAs have been produced for different species as their genome sequence has become available. The process of building the SSPAs has been carried out in the MicroScope platform. A total of 19 SSPAs have been produced in the course of the project (the genome sequences of the remaining species on the Microme target list have not yet been generated). The SSPAs have been represented in BioPax file format and are available for download, as part of the version 4.0 data release, at http://www.microme.eu/download-microme-data.

Table 2. SSPAs included in the Microme data set 4.0

Acinetobacter sp. ADP1
Bacillus subtilis subsp. subtilis str. 168
Beutenbergia cavernae DSM 12333
Brachybacterium faecium DSM 4810
Chitinophaga pinensis DSM 2588

Delftia acidovorans SPH-1
Dyadobacter fermentans DSM 18053
Escherichia coli str. K-12 substr. MG1655
Kytococcus sedentarius DSM 20547
Pedobacter heparinus (strain ATCC 13125 / DSM 2366 / NCIB 9290)
Pseudomonas aeruginosa (strain ATCC 15692 / PAO1 / 1C / PRS 101 / LMG 12228)
Pseudomonas putida KT2440
Rhizobium etli (strain CFN 42 / ATCC 51251)
Rhodopirellula baltica (strain SH1)
Salmonella enterica subsp. enterica serovar Typhimurium str. SL1344
Spirosoma linguale DSM 74
Tsukamurella paurometabola (strain ATCC 8368 / DSM 20162 / JCM 10117 / NBRC 16120 / NCTC 13040)
Xylanimonas cellulosilytica DSM 15894
Roseobacter litoralis (strain ATCC 49566 / DSM 6996 / JCM 21268 / NBRC 15278 / OCh 149)

Milestones Achieved

Milestone M9: Release of Microme pathways dataset 4.0 including 20 curated SSPAs

The data set has been released included 19 curated SSPAs (compared with the expected number of 20). The infrastructure for the generation of SSPAs is in place, and production of further SSPAs will continue as the necessary genome sequences become available in future.

Perspectives

The data collection and inference pipelines are now in maintaining mode. It is planned to keep these operating beyond the formal life-end of the project and to make the data available through microme.eu and other EBI-operated databases.