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**

D2.6 Microme data set release 3.0

Due date of deliverable: 30.11.2012
Actual submission date: 30.11.2012

Start date of project: 1.12.2009 Duration: 48 months

Organisation name of lead contractor for this deliverable: EMBL-EBI

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 deliverable has been prepared by the EBI. The coordination of release generation is led by EBI. Other partners making significant contributions to the process include CEA (Microscope annotation and platform provision), SIB (reaction/chemical entity curation, UniProtKB curation) and AMB (*E. coli* re-annotation).

INTRODUCTION

D2.6: Microme data set release 3.0

The aim of this deliverable was to produce a third Microme data set, enhancing the reactions and pathways prepared as part of the second Microme release (see deliverable D2.5).

Methods

The preparation of this deliverable involved the following work:

(i) development of software for the import of content provided by external sources and
(ii) curation activities designed to plug the gaps in the inferable metabolic networks;
(iii)design and implementation of a robust deployment strategy;
(iv)statistical analyses including data mining for reporting and multivariate exploratory data analyses;
(v) Reorganisation of the Microme public FTP site providing access to sequential releases of the data set.

Results (if applicable, interactions with other workpackages)

The third Microme data release has integrated data from nine inter-related sources. A total of 1,911,029 reaction instances have been inferred over 4,608 genomes, representing components of 1928 pathways. Pathways are defined in EcoCyc[1] (release 15.0) and MetaCyc[2] (release 16.0). The sources of the genome data are submissions made to the International Nucleotide Sequence Databases. The sources of the reaction data are three sources of externally curated or predicted data, and three methods of automatic reaction inference (whose development is described in deliverables D1.4 and D1.8). These six methods are:

(i) enzymatic function curated in the UniProt Knowledgebase[3].
(ii) reactions imported from EcoCyc.
(iii) curated and predicted genome-protein-reaction associations for 874 species imported from the Microscope[4] platform.
(iv) automatically inferred InterPro[5] annotations of protein sequences, indicating the presence of protein domains, the curated associations of InterPro entries and Gene Ontology[6] (GO) terms, indicating protein functions, and curated associations of Gene Ontology functions to enzymatic functions, as defined by the Enzyme Commission[7], and reactions in RhEA[8].
(v) projections of the EcoCyc data to orthologous proteins identified by Microme through the use of the Ensembl Compara platform[9].
(vi) protein enzyme-reaction associations, as defined by the Enzyme Commission, predicted using the PRIAM[10] methodology.

This data has been assembled using the projection strategies developed in WP1 (see D1.8), and additionally supplemented by curation activities designed to plug the gaps in the inferrable networks.

Curation Activities for Data Release 3.0

Curation of InterPro entries

In order to improve the accuracy and coverage of the inference methodology used in the Genome-Reaction Matrix, we reviewed 1,211 InterPro entries related to transporters and 2,621 entries related to proteins involved in metabolism, by extracting evidence from published experimental data. According to the InterPro approach to GO annotation, GO terms assigned to a given entry need to represent ~95% of its members (PMID: 22301074). On this basis, 420 InterPro entries were associated with new or improved (i.e. more specific) GO terms.

GO-Rhea mappings

All reviewed InterPro entries (3,832) that had an associated term in the GO Molecular Function ontology (assigned either by Microme curators, or by InterPro curators in previous work.) were again checked to ensure that all terms that could be translated into a Rhea reaction were duly cross-referenced, resulting in 113 new GO-Rhea mappings.

Curation of new Rhea/ChEBI entities

During this curation work, new GO terms, 114 new reactions were curated in the Rhea database, and, where necessary, associated chemical entities were created in the ChEBI database. ChEBI serves as a reference repository for chemical entities in Microme, and Rhea as the reference repository for reactions.

Impact

Table 1 shows the impact of the curation efforts to date on the size of the genome reaction matrix. Targeting one database – InterPro – for enrichment has had have positive knock-on effects on other resources and a cumulative snowball effect on annotation, because of the possibility to automatically apply this information to thousands of genomes.

Table 1: The impact of curation efforts on the Microme Genome-Reaction matrix.

	Total	***Affected***	***Perc (%)***
Genomes	*4149*	*4134*	*99.64*
Gene-reaction associations	*3726498*	*303737*	*8.28*
Genome-reaction associatons	*1729154*	*141945*	*8.21*
Avg reactions/genome	*416.9*	*34*	*8.16*

Re-annotation of the genome of Escherichia coli K-12 substr. MG1655 using the MicroScope platform

A key goal of Microme is to generate well-annotated reference data for a number of species, which can be used to reconstruct metabolism in a large number of other organisms. The update of the annotations of the genome of *Bacillus subtilis subsp. subtilis str. 168* (by AMB, CEA, SIB and CERTH) has been reported in a previous deliverable (D2.). The Firmicutes clade is far from representing bacteria in general, and so, more reference species need to be (re-)annotated (e.g., *Pseudomonas putida KT2440* and *Salmonella enterica subsp. enterica serovar Typhimurium str. SL1344* are two species targeted for particular attention by Microme). However, there is currently no single consistent annotation even for the model bacterium *Escherichia coli*. Even combining several resources (EcoGene, EcoCyc, RefSeq, and UniProt), the present annotations lack information that could allow the prediction of various gene functions (in particular related to metabolism). To overcome this limitation, we therefore decided to carry out the re-annotation of the genome of E. coli K-12 substr. MG1655 using the MicroScope platform. This will assist in the later annotation of other closely related species.

MicroScope (https://www.genoscope.cns.fr/agc/microscope/home/index.php), the Microbial Genome Annotation & Analysis Platform, is a web-based platform for microbial comparative genome analysis and manual functional annotation. The MicroScope web interface, MaGe (Magnifying Genomes), was designed to assist curators in the evaluation of all available data needed for assigning the best possible annotation to a given gene product.

An automatic annotation of the genome of *E. coli K-12 substr. MG1655 in* the MicroScope platform has been generated by the CEA, and the AMB and the EBI have now begun to validate these annotations. This curation work consists of, but is not limited to, checking and updating, if necessary, the gene product and product type, adding synonyms, EC numbers, MetaCyc and Rhea reactions, cellular localization, biological process, roles, documenting a comprehensive and up to date list of PubMed references, and, finally, adding relevant information in the comments field. To date, we have manually re-annotated approximately 25% of the genes. Activities are ongoing.

The data

This new version of the Microme data set significantly extends the content of release 2.0. We put particular attention in selecting and integrating additional data with the main goals of increasing the coverage of the taxonomical space and the quality of the inferable draft metabolic networks. 1870 new genomes from the International Nucleotide Sequence Database have been imported and the six methods outlined above, where applicable, applied to all new and pre-existing genomes. The overall effect on the data set is summarised in Table 2.

	Release 2.0	**Release 3.0**
Genomes	2738	4608
Coding sequences	1371574	3466149
Unique RhEA reactions	1489	2607
Unique gene-reaction associations	1809974	4129454
Unique genome-reaction associations	788588	1911029
Avg/Max reactions per genome	288.1/594	414.9/1861
Avg/Max pathways per genome	162.9/293	664.5/1152

Table 2. Comparison between current (v3.0) and previous (v2.0) Microme data set releases.

Below are a number of figures that outline the composition of the data set in more detail.

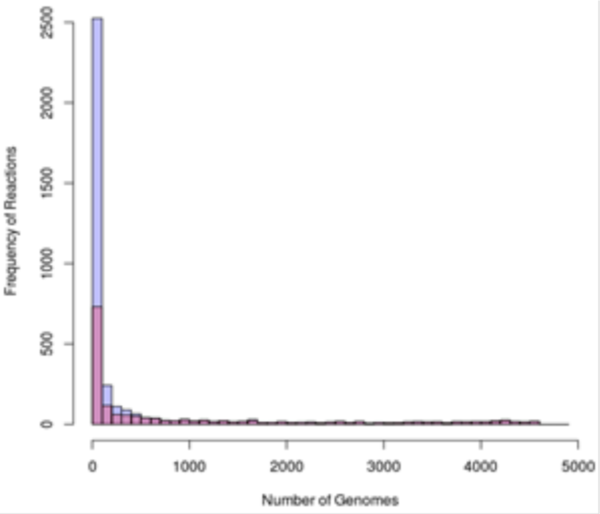

Figure 1. Histogram of the number of reactions per genome.

Figure 1 shows the histogram of the number of reactions per genome, with the purple bars referring to the statistics restricted to the reactions inferred from the method borrowed from the previous release, and based on InterPro annotations. Overall, fewer genomes containing larger number of reactions with the extreme cases to the left of the diagram correspond to data extracted from the MicroScope platform, in which a small number of genomes have been exceptionally well-annotated.

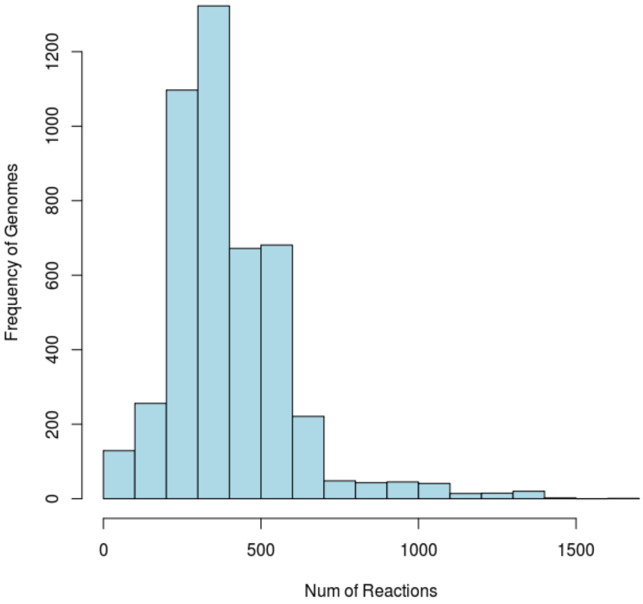

Figure 2. Histogram of the number of genomes per reaction in which each reaction is annotated

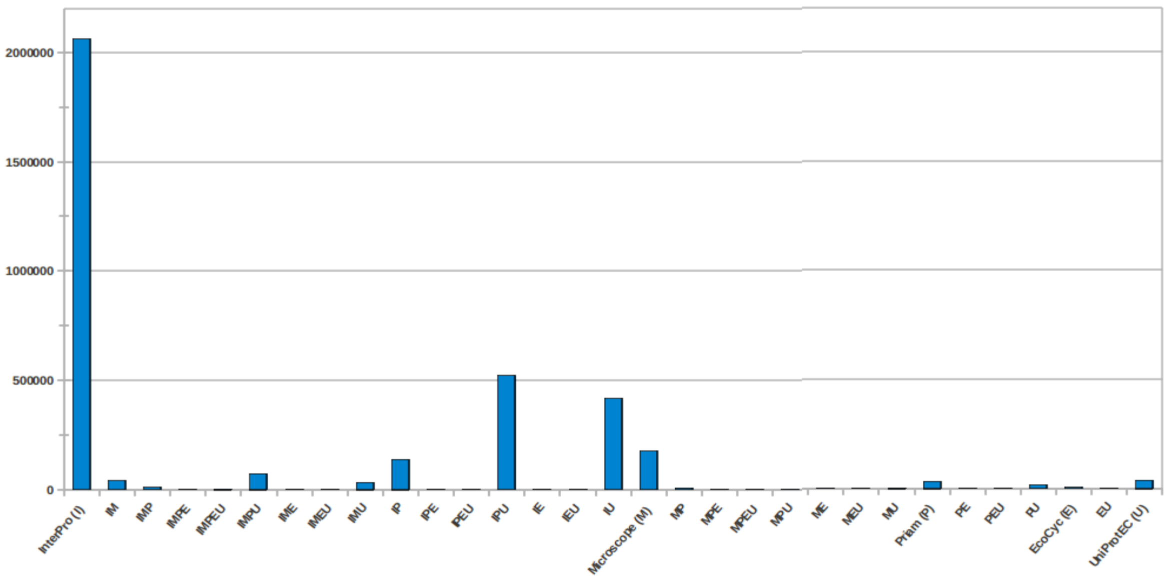

Figure 3. Histogram of the number of unique gene-reaction associations predicted per (combination of) method(s). Each method is represented with the initial letter. Each bar represents the number of gene-reaction associations predicted together by the combination of methods identified by the corresponding 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 so on.

The InterPro method alone amounts at a large fraction of the total number of unique GPRs, with significantly less contribution from the other methods. The predictions of the methods based on InterPro, PRIAM and direct UniProt-EC associations largely intersect and represent the second largest contribution, followed by InterPro and UniProt-EC together. This is a consequence of the ability to automatically apply InterPro classification to many genomes, and is the reason for concentrating curation efforts on increasing the number of reactions that can be inferred by this method.

<u>Statistical analyses</u>

We have performed multi-variate data analysis by applying PCA to the matrix of reactions inferred over the total number of genomes. Figure 4 shows the plot of the projections of the genome in the subspace spanned by the first two principal components (left) and the coefficients of the variables (i.e. reactions) in the same subspace (right). The latter plot reveals how PCA is able to identify a certain degree of correlation among the reactions and how this clusters along the main directions of variance. The former plot indicates similar clustering patterns of the genomes. We're currently investigating in more detail the structure of the data based on these preliminary results.

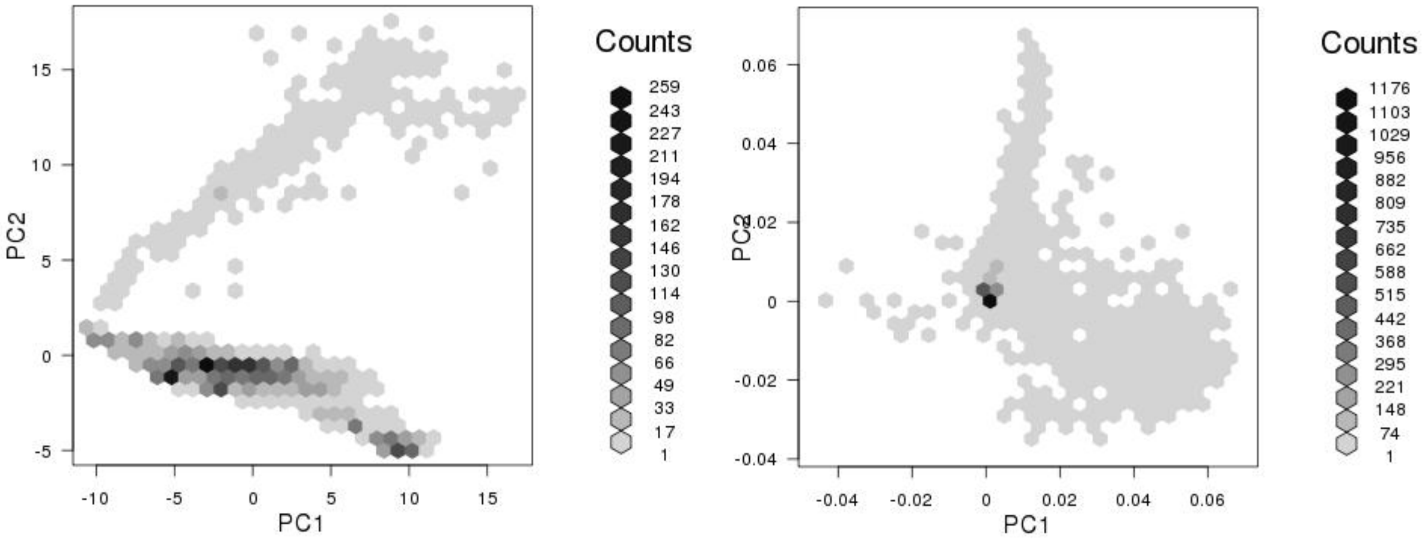

Figure 4. PCA analysis: scoring (left) and loadings (right) plot.

Availability

Data has been made available through the Microme genome matrix browser (http://bacteria.microme.eu/matrix/index.html), and through the Microme FTP site (ftp://ftp.ebi.ac.uk/pub/databases/microme/releases/current/reaction_matrix)

URLs

1	http://www.ecocyc.org
2	http://www.metacyc.org
3	http://www.uniprot.org
4	http://www.genoscope.cns.fr/agc/microscope
5	http://www.ebi.ac.uk/interpro
6	http://www.geneontology.org
7	http://www.chem.qmul.ac.uk/iubmb/enzyme
8	http://www.ebi.ac.uk/rhea
9	http://www.ensembl.org/info/docs/compara/index.html
10	http://priam.prabi.fr

Perspectives

We are continuing to integrate new methods for prediction, to incorporate new curated data, and to make new releases at monthly intervals including additional genomes as they are submitted to the international nucleotide archives. We continue to investigate the structure of the data (decreasing overlap of predictions/increasing orthogonality of predictions). Future work will also involve the incorporation of data resulting from the experimental analysis of *Salmonella* currently being undertaken by beneficiary WTSI.