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**

D2.5
Microme dataset release 2.0

Due date of deliverable: 1/11/2011
Actual submission date: 17/02/2012

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 EBI. The data set is the product of combined effort by Amabiotics, CEA, EBI, SIB, and WTSI.

INTRODUCTION

D2.5: Microme dataset release 2.0

The aim of this deliverable was to produce a second Microme data set, enhancing the reactions and pathways prepared as part of the first Microme release (deliverable D2.3).

Methods

The preparation of this deliverable involved the following work:
(i) Analysis of genome annotation, annotation tools, and associations between different data types present in public database resources.
(ii) Development and QC of code to extract and integrate all reaction/pathway/genome associations present in the public database resources.
(iii) Analysis of the reaction networks inferable across microbial species, and in individual species.
(iv) Development and population of a relational database to support access to the data, and future development of interactive interfaces.
(v) Development of a flat file format for bulk access to the data.
(vi) Publication of the data set on the Microme portal.
(vii) Provision of public access to a mysql server providing access to the data.

Results (if applicable, interactions with other workpackages)

To prepare the second data release for Microme, we prepared a reaction matrix over 2707 complete bacterial and archaeal genomes. The source for this data is (i) genome sequence and protein-coding gene annotations submitted to the international nucleotide archives (ii) InterPro annotations of the protein sequences (iii) curated associations of InterPro entries (indicating protein domains) and gene ontology terms (indicating protein functions) and (iv) curated associations of Gene Ontology (GO) functions to reactions (as defined in the RhEA database).

By combing these data sets, it is possible to produce a matrix of reactions whose presence can be inferred in an annotated genome. Code was prepared to combine these data which can be rapidly run to update the matrix to accommodate new genomes, new InterPro methods and new associations (via GO) between InterPro entries and reactions. Because InterPro classification of genomes is a computational (and not a curational) process, this is a scalable method. While it does not provide a complete classification of all reactions, it does provide a basic template of reactions which can be supplemented by more detailed knowledge from particular species.

The approach does not take account of complex structure. The InterPro-GO and GO-RhEA assignements are

made on the basis that the presence of the InterPro domain or family implies the that the GO term applies to the product; the definition of some GO terms in the "Molecular Function" ontology could be paraphrased as "catalyst of reaction X" and in these cases, a cross-reference to RhEA is maintained. Because such terms are usually applied to individual proteins, the annotation may not reflect the structure of complexes. Specifically, the method will not detect the non-presence of reactions whose catalytic subunits are present but whose non-catalytic subunits are absent. Because of possible incompleteness and errors in annotation, the partial coverage of the protein space by InterPro, and incompleteness in the InterPro-GO and GO-RhEA assignments, the resulting matrix is necessarily incomplete. However, it represents the total knowledge inferable from the extant annotation in the various sources.

Some summary data for the process is given in table 1.

Table 1: Summary of Data Used to Generate Reaction Matrices for 2707 Genomes

Number of bacterial/archaeal genomes (ENA release 110)	8388
Number of protein coding annotations	18437611
Average numebr of protein coding annotations/genome	2198
Number of InterPro entries (InterPro release 35.0)	22361
Number of InterPro entries with mapping to GO	10332
Number of GO terms (GO release 1.2614)	35791
Number of GO terms with mapping to RhEA	2600
Number of protein sequences with mapping to RhEA	1371574
Total number of reactions (RhEA release 28)	17895
Number of reactions detectable in at least 1 bacterial/archaeal genome	1489
Number of genomes with at least 1 detectable reaction	2707
Total number of genome-protein-reaction instances	1853662
Total number of genome-reaction instances	782022
Average number of reactions/genome (genomes where number of reactions > 1)	289

For this data release, we enhanced the default matrix with the set of additional reactions incorporated for *Escherichia coli* from the EcoCyc resource and the iAF1260 metabolic model[1] that were used to generate the initial data release. This wider set of reactions had previously been projected (using orthology) to 30 key species that have been identified as priority species for Microme.

[1] A M Feist, C S Henry, J L Reed, M Krummenacker, A R Joyce, P D Karp, L J Broadbelt, V Hatzimanikatis & B Ø Palsson. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information, ***Molecular Systems Biology* 3:121**, 2007

We have also converted the reaction matrix into a skeleton metabolic network for each species, indicating potential routes of metabolic flux. Reactions with common metabolites (among their inputs/outputs) have been linked; selected small molecules with high connectivity in the network (e.g. water) are assumed not to be the major sources of flux in the system and have been filtered from the network. The network has been visualised with the Cytoscape graph visualisation package, to check for its coherence. One can visualise the full network of RhEA reactions; the network of RhEA reactions that could possibly have been inferred using this approach (i.e. reactions that can be inferred through classification by InterPro, and subsequently through the assignment of GO terms); and the inferred network in particular species. Overall properties of one of these networks are given in table 2. An example of comparison of inferred reaction networks for two species (E. coli, B. subtilis) is shown in figure 1.

The networks from selected different species have been co-visualised and to highlight common and unique reactions. New user visualisation tools are currently in development to allow users to explore this data through the Microme website (work package 6). The data set will also be used as the basis for the construction of stoichiometric models (work package 3).

Table 2: Properties of Selected Reaction Networks

Network	Nodes (i.e reactions)	Edges	Connected Subgraphs	Isolated Nodes	Avg Neighbours	Network Diameter
All Species	1489	21871	151	130	29.4	13
E. coli	577	2387	80	66	8.3	12
B. subtilis	432	1279	73	56	5.9	14

To further analyse this data, we have clustered genomes according to their reaction profile, and reactions according to their species distribution, using R library Heathmap2[2], scoring each reaction for either presence (value 1) or absence (value 0). Reactions have been clustered according to their similarity of distribution, and species clustered according to the similarity of the pathways they contain.

[2] http://rss.acs.unt.edu/Rdoc/library/Heatplus/html/heatmap_2.html

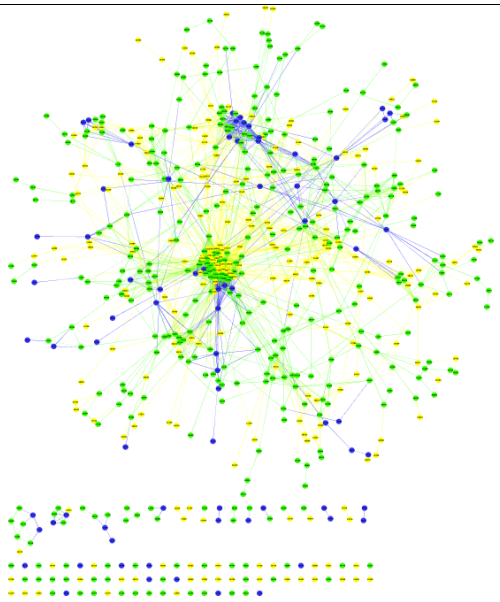

Figure 1. Comparison of inferred reaction networks for two species. Nodes in blue (resp. yellow) refer to reactions occurring only in E. coli (resp. B. subtilis); nodes in green represent reactions occurring in both species.

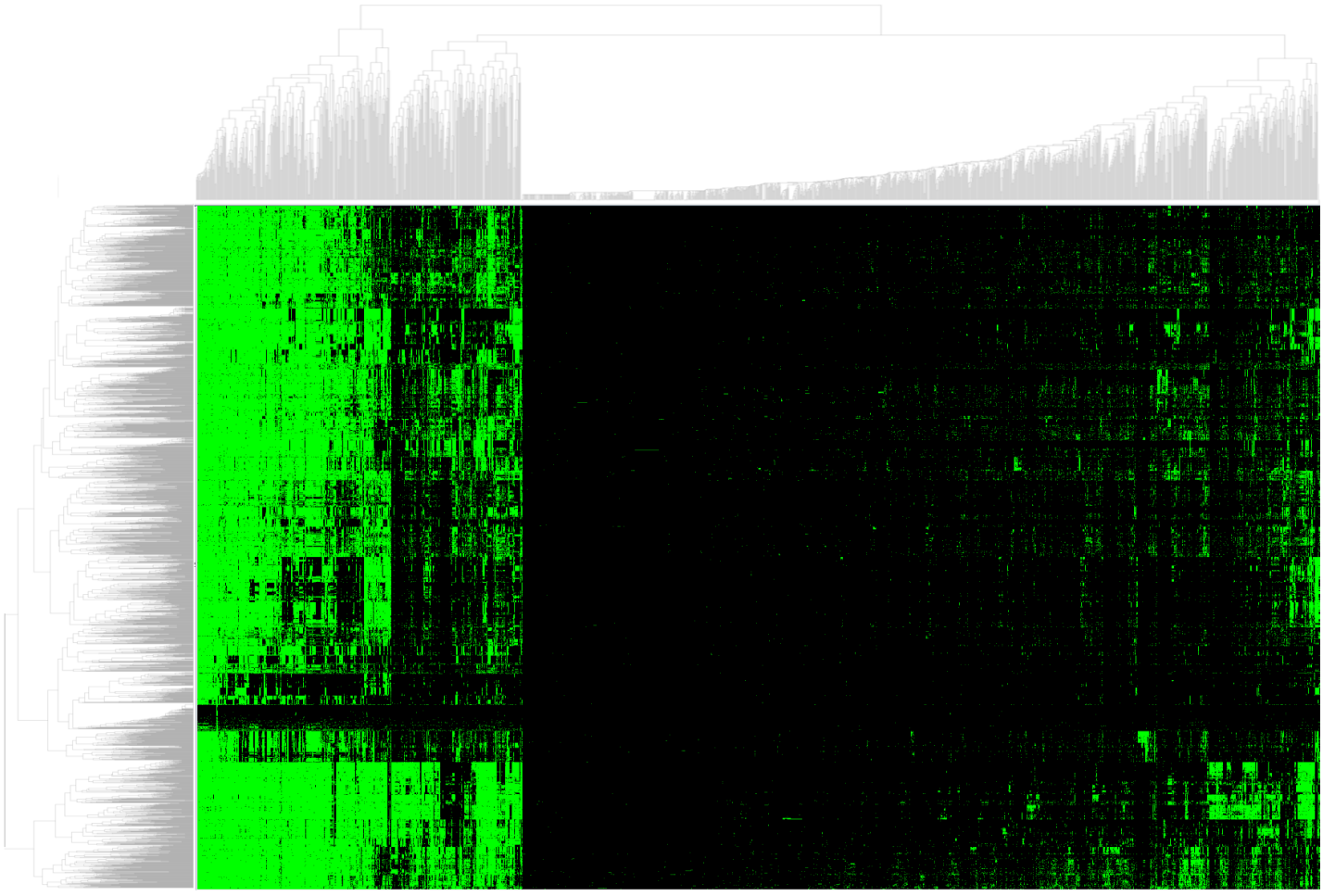

Figure 2. Reaction-Genome matrix clustered with heatmap2 library in R

The reaction matrix has additionally been mapped onto known pathways from EcoCyc. Each pathway has been assessed for completeness and scored according to the percentage of reactions present in the matrix. Pathways have been clustered according to their similarity of distribution, and species clustered according to

Perspectives
With this deliverable, the set of reference pathways and reactions previously prepared have been supplemented (and used to interpret) the full matrix of reactions inferable from the public annotation of microbial genomes. The Microme portal (see WP6) will be developed to provide new interactive interfaces to this data. The data sets will be used in WP3 for the preparation of draft metabolic models. In WP2, the data set will be updated at regular intervals, and integrated with new pathway and reaction data curated and imported from other sources.