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**

Deliverable D7.1
First Microme Annotation Jamboree

Due date of deliverable: M22
Actual submission date: M20

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)	

Revision [1.0]

Contributors

Responsible Beneficiary: CEA
Lead and coordinated by CEA
Staff initially involved in DOW: EMBL-EBI, CERTH, CNIO, SIB, ULB
Additional staff involved in the jamboree annotation: WTSI, MN, KO and TAU.

INTRODUCTION

Deliverable reference number: D7.1, First Microme Annotation Jamboree

This document summarizes the work done during the first Microme Annotation Jamboree that took place at the Genoscope (CEA) during the 28th of June 2011. The Jamboree focused on the curation of Pseudomonas putida KT2440 metabolic pathways based on Microme data and tools currently available from different members of the consortia.

ORGANISATION

Location : CEA - Institut de génomique 2, rue Gaston Crémieux CP5706 91057 Evry CEDEX France.
http://www.genoscope.cns.fr/spip/spip.php?lang=en

Number of participants : 24 (3 from EBI + 2 EBI teleconf, 2 ULB, 2 WTSI, 2 MN, 1 CSIC, 1 WUR, 1 SIB, 2 AMB, 6 CEA, 2 TAU).

This Jamboree has followed a training day (27th June 2011) during which the set of tools used (i) to produced the data to be curated and (ii) to guide the curation process and to enter reaction/pathway into Microme were presented.

Agenda:

I. Microme-Reactome exploration interface: How can curators explore the metabolic data available in Microme ? – How can curators enter new records, modify existing records, add cross-references ?

=> Practical curation exercice using Microme data and one "case study" proposed by CEA/SIB preliminary analysis

II. Workshop *P. putida* "Metabolic reconstruction"

Goals: assess if the predicted pathways are present or not (false positive) + completion of the pathways where some reactions are missing
Starting data: Results of the Microme and MicroScope (Pathologic) projections were compiled in a worksheet to guide the curation process. Attendees worked on specific set of pathways for which curation was needed.

Organisation: organisation of the workshop – practical work – restitution of the curation process

III. Workshop "Genomic centric curation": following selected pathways of interest in *P. putida*

Goal: Annotation of some metabolic pathways for which we have BIOLOG results and which are not in KEGG and METACYC (if possible).
=> Each Biolog experiment indicates whether the organism is able to grow with a given compound as sole carbon source. A predicted pathway can be evaluated by checking that the initial substrate and intermediate compounds can be used as carbon source by the considered organism.

Starting data: We selected a few number of interesting *P. putida* metabolic pathways (combining projection, pathway discovery and Biolog results).

Organisation: organisation of the workshop– practical work – restitution of the curation process

DATA

Reference data used during the Jamboree are described in the following tables.

Table 1: Microme Central Repository statistics.

Data	*Numbers*	*Shared with Reference Data set*	*Shared with MetaCyc Data set*
small molecules	*1386*	*ChEBI:1100*	*676*
reactions	*2352*	*RHEA:1266*	*2180*
pathways	*316*		*316*
reference organisms	*1*	*NCBI:1 (id 511145)*	
reference genes	*1869*	*Ens. Bact.: 1733*	*1776*
reference proteins	*1766 polypeptides 771 complexes*	*Uniprot: 1740*	*1724*
reference activity	*1418*		*1346*

Table 2:Microme Projected Data Repository

Organisms	*24*
Genes	*111116*
Ortholog groups	*10998*
Orphan genes	*20965*
Projected Activities	*22711*
Not projected reactions	*32855*

All materials (computation of orthologs using Compara tool, results of the projection, results of the metabolic network comparisons, etc) can be found on the Microme wiki (http://www.ebi.ac.uk/seqdb/confluence/display/Microme/WP7).

METHODS. Software tools used during the Jamboree for pathway curation.

- *MICROSCOPE (CEA):*
  Computational platform for the annotation and comparative analysis of bacterial genomes. In the context of the Jamboree, novel functionalities have been added to the MICROSCOPE platform in order to support curation at the level of gene or pathway; this includes:

- **The "MetaCyc Reaction" field** in the gene editor page of MICROSCOPE, which allows the user to link one or more metabolic reactions from MetaCyc database to each gene of the genome. A search tool has also been added to query MetaCyc reactions associated to a particular EC number or a particular keyword (tutorial : https://www.genoscope.cns.fr/agc/website/spip.php?article593)

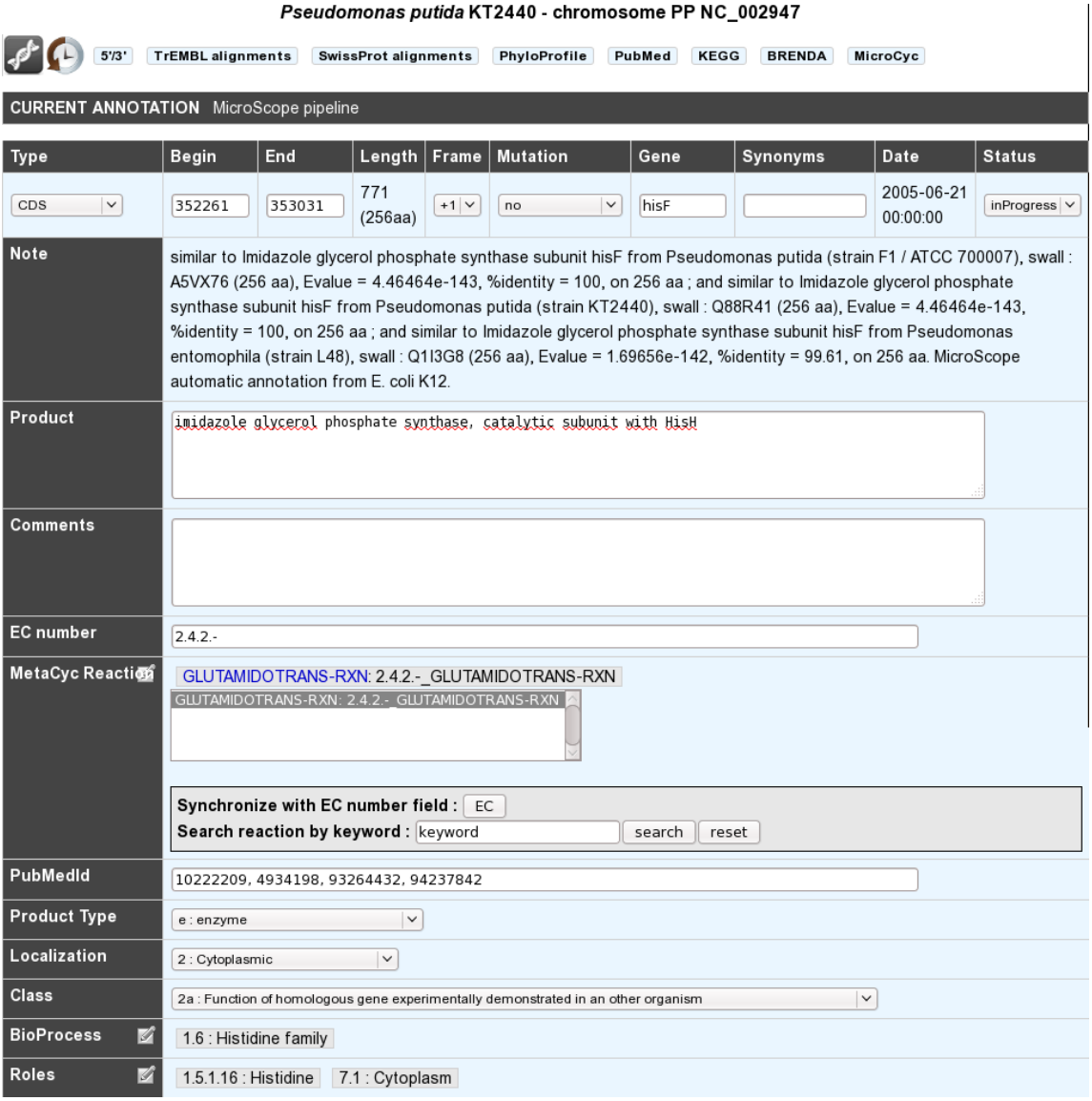

Figure 1: Each gene can now be linked to a catalytic activity based on its EC numbers but also directly to a MetaCyc reaction. This feature ensures that Microscope is fully compliant with chemical information. The curator is able to search for a MetaCyc reaction based on keywords (EC numbers, catalytic activity name and substrate/product names). The pathway projection algorithm (Pathologic) is launches every night to take into account newly curated data if any.

- The **"Metabolic profiles" functionality**, which allows to compare the metabolic profiles of several genomes based on the degree of completeness of their metabolic pathways in BioCyc or KEGG resources. The corresponding results directly highlight pathway holes in each selected organism and allow one to make cross-species comparison. (tutorial : https://www.genoscope.cns.fr/agc/website/spip.php?article624)

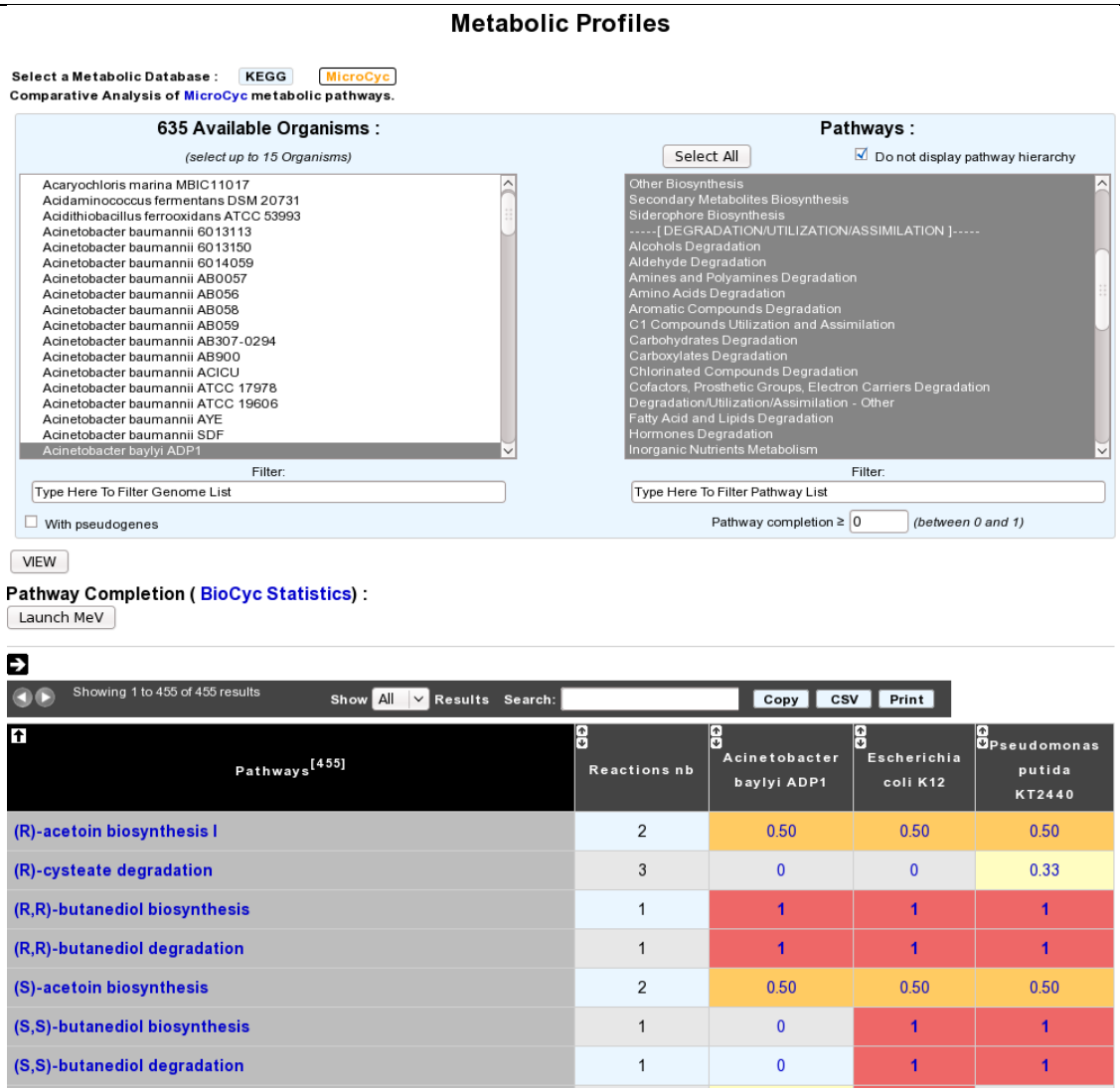

Figure 2: The Metabolic Profile allows to select several organisms and the pathways of interest. The tool will compute a table where each row is associated to a given pathway (either in the BioCyc or KEGG hierarchy): the first column gives the total number of reactions in the pathway, the following columns give the result of the 'pathway completion' value in each compared organism, i.e, the number of predicted reactions in the pathway divided by the total number of reactions in the same pathway.

- *MICROME-REACTOME curator tool (EBI):* Software tool that allows the management of Microme reference data on microbial metabolism available in the pre-release dataset (chemical compounds, reactions, pathways, Gene-Protein-Reaction associations). In the context of the Jamboree, the MICROME-REACTOME curator tool was used for "de novo" creation of metabolic pathways from Microme reference data: a complete tutorial can be found at http://wiki.reactome.org/index.php/New_Reactome_Curator_Guide.

RESULTS: Curated data

Pathway design in Microme environment using the Reactome curation tool (part I of the Agenda)
A « case study » for de novo creation of a metabolic pathway not present in public resources has been selected : the pathway for UDP-N-acetyl-L-fucosamine. UDP-N-acetyl-L-fucosamine is a precursor of L-fucosamine, one of the components of the lipopolysaccharide structure of *Pseudomonas aeruginosa*, and its complete biosynthetic pathway from nucleotide sugar precursor UDP-N-acetyl-D-glucosamine has been characterized in Mulrooney et-al 2005:

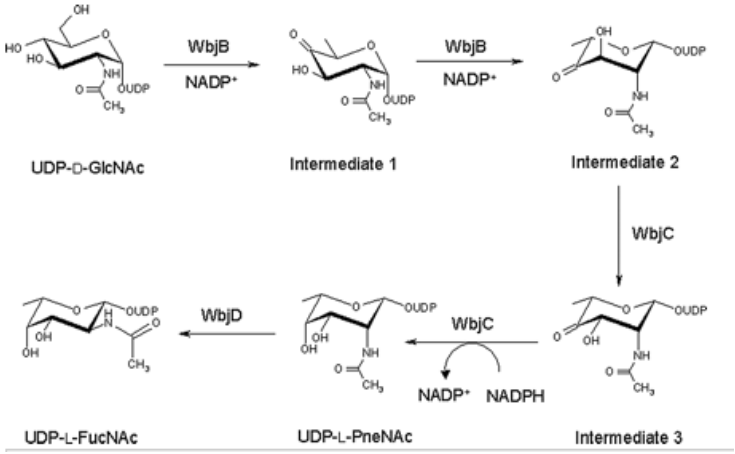

Figure 3: UDP-N-acetyl-L-fucosamine biosynthesis pathway in *Pseudomonas aeruginosa* serotype O11. Adapted from Mulrooney et-al 2005.

This pathway is not described in reference pathway resources such as METACYC, KEGG, UNIPATHWAY. In order to create this metabolic pathway in Microme, all compounds and reactions should first be present in the reference resources that is, in CheBi for the compounds and in Rhea for the reactions:

- **Compounds:** Only UDP-D-GlcNAc (CHEBI:57705), intermediate 2 (UDP-2-acetamido-2,6-dideoxy-ß-L-arabino-hex-4-ulose(2-) (CHEBI:60101), and UDP-L-FucNAc (CHEBI:62241) are present in CHEBI, i.e correct structure at ph 7.3 (needed for RHEA reaction creation). The rest of the compounds of the metabolic pathway has been created in ChEBI using the ChEBI submission tool available at https://www.ebi.ac.uk/chebi/submissions/login. Indeed, the following compounds have been created:

Table 3: Compounds created for the jamboree.

Compound	*ChEBI name*	*CHEBI ID*
Intermediate 1	*UDP-2-acetamido-2,6-dideoxy-a-D-xylo-hex-4-ulose(2-)*	*CHEBI:62375*
Intermediate 3	*UDP-2-acetamido-2,6-dideoxy-ß-L-lyxo-hex-4-ulose(2-)*	*CHEBI:62377*
UDP-L-PneNAc	*UDP-2-acetamido-2,6-dideoxy-ß-L-talose(2-)*	*CHEBI:62372*

- **Reactions:** when compounds were absents in ChEBI the corresponding reactions were also absent in RHEA. Once created, compounds were reviewed by ChEBI curators (« Checked » status in the database); the corresponding reactions were then created in RHEA following the above diagram:

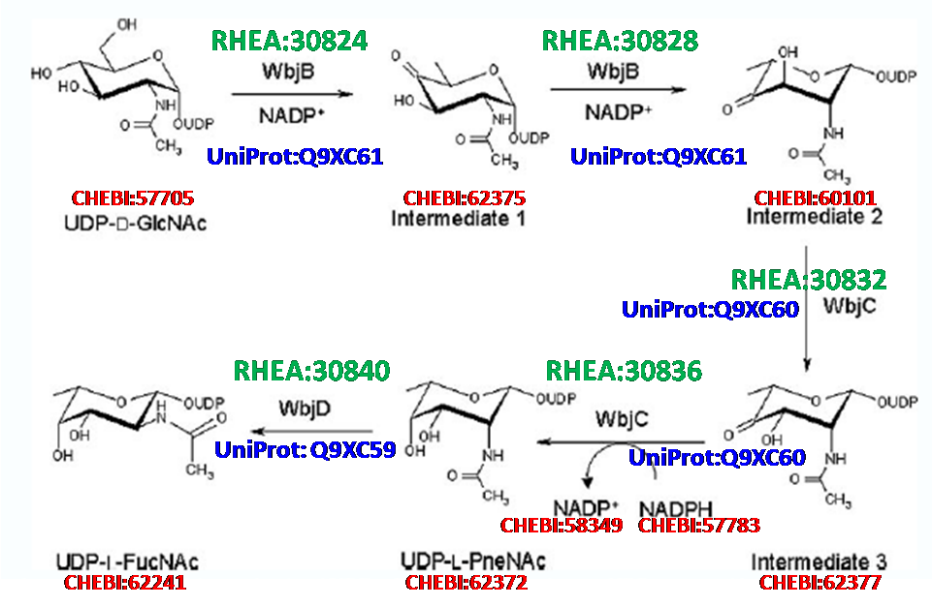

Figure 4: Final state of the projection of compounds and reactions of UDP-L-FucNAc biosynthesis pathway in ChEBI and RHEA respectively.

Finally, the UDP-N-acetyl-L-fucosamine biosynthesis pathway diagram was created in Microme environment. For this purpose, the Reactome curator tool was used : it allows *de novo* creation of pathway diagrams locally based on reference data available on Microme repository (compounds, reactions, gene associations).

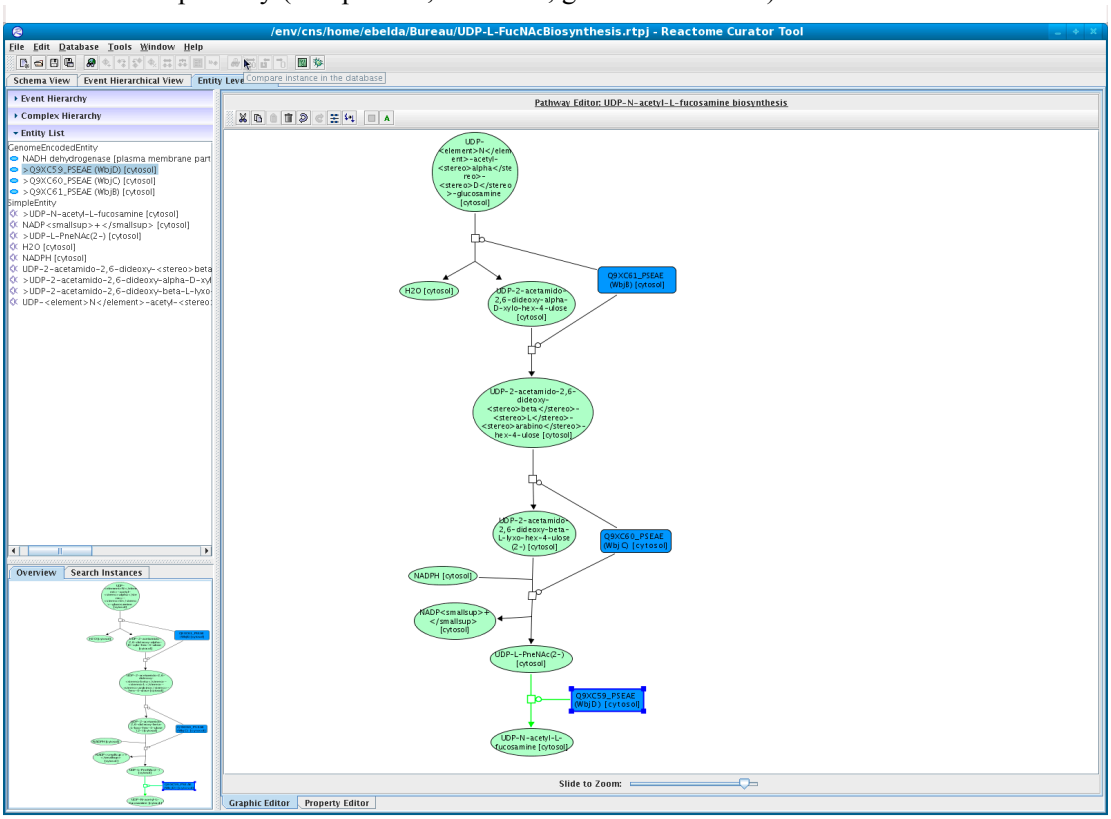

Figure 5 : Snapshot of the UDP-N-acetyl-L-fucosamine biosynthesis pathway created with the Reactome curator tool. Green elements represents metabolites defined based on newly created ChEBI elements. Connections between metabolites represents reactions created based on newly created RHEA reactions. Blue rectangles represents associated catalyst of each reaction step.

Pathway validation in *Pseudomonas putida* KT2440 into Microscope (part II and III of the Agenda)

The *Pseudomonas putida* KT2440 metabolic network was reconstructed using the MicroCyc component of MicroScope. Gene-Protein-Reaction associations were based on blastP similarity searches using reference genomes (*Escherichia coli* K12 - EcoCyc annotation and *Acinetobacter baylyi* – CEA bacterial model) and the Bidirectional Best Hits criteria. These associations were projected on the MetaCyc pathways using the BioCyc PathoLogic software and MetaCyc as reference pathway repository. We obtained:

- 1865 GPR associations made of 1301 distinct genes and 1122 distinct reactions
- 307 pathways were predicted, of which 171 were complete
- Among the 136 incomplete pathways, 267 gap reactions were detected (i.e. reactions for which no gene in the organism codes the enzyme)

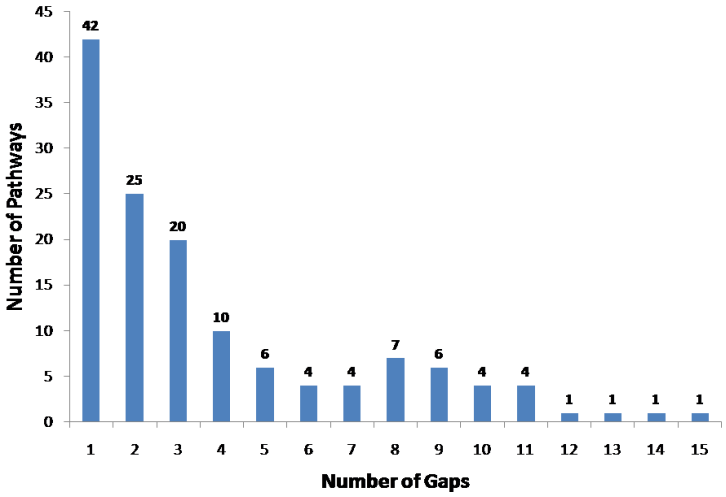

Figure 6: Distribution of gaps among pathways: 171 of the predicted pathways are complete and 236 have at least one missing activity.

Some projected metabolic pathways were selected in the context of the curation process :

Pathways	Reactions nb	Acinetobacter baylyi ADP1	Escherichia coli K12	Pseudomonas aeruginosa PAO1	Pseudomonas putida KT2440
cis-dodecenoyl biosynthesis	6	0.33	1	1	0.83
isoleucine degradation I	6	0.50	0.33	0.67	0.50
leucine degradation I	6	0.33	0.33	0.67	0.83
threonine degradation I	4	0.75	1	0.50	0.50
lysine degradation IV	5	0.20	0	0	0.80
valine degradation I	8	0.50	0.13	0.75	0.63
tyrosine degradation I	5	0.20	0	1	0.80
quinate degradation I	3	1	0.33	0.33	0.67

Pathway	*MicroCyc Link*	*Number of reactions*	*Number of missing activities*
Cis-dodecenoyl biosynthesis	*http://tinyurl.com/metacyc-dodecenoyl*	*6*	*1*
Leucine degradation	*http://tinyurl.com/metacyc-leucine*	*6*	*1*
Lysine degradation	*http://tinyurl.com/metacyc-lysine*	*5*	*1*
Threonine degradation	*http://tinyurl.com/microcyc-threonine*	*4*	*2*
Tyrosine degradation	*http://tinyurl.com/microcyc-tyrosine*	*5*	*1*
Isoleucine degradation	*http://tinyurl.com/microcyc-isoleucine*	*6*	*3*
Valine degradation	*http://tinyurl.com/microcyc-valine*	*8*	*3*
Quinate degradation	*http://tinyurl.com/microcyc-quinate*	*3*	*1*

For these 8 metabolic pathways, Microscope pathway projections yielded incomplete pathways with one or two missing reaction(s) step(s). Using the tools available on the Microscope platform, it was possible to retrieve the missing enzymatic activities in *Pseudomonas putida* KT2440 and to add the corresponding Gene-Protein-Reaction associations in the gene editors.
During the workshop sessions attendees were grouped by 2, and together they curated two out of the nine proposed metabolic pathways. Microscope support team were available to curators in order to help them in the use of the different functionalities of the Microscope platform**.** The results of the curation for these proposed pathways are available on the confluence page (http://www.ebi.ac.uk/seqdb/confluence/display/Microme/WP7, *PathwaysJamboreeWorkshop.pptx*)

To illustrate an example of the curation process using the MicroScope functionalities, we selected a more complicated example: the **L-carnitine degradation pathway**. Pseudomonas species can assimilate and utilize L-carnitine as the unique source of both carbon and nitrogen. Accordingly, Biolog data from DSMZ partner shows the capability of *Pseudomonas putida* KT2440 to grow with the L and D-carnitine substrate.

Table 4: Growth phenotype profile for D,L carnithine

Compound	*CHEBI ID*	*Biocyc ID*	*BIOLOG Acinetobacter Baylii*	*BIOLOG Paeruginosa*	***BIOLOG Pputida***	*BIOLOG Ecoli*	*BIOLOG Bsubtilis*	*BIOLOG Salmonella*
D,L carnitine	*17216*	*L-CARNITINE*	*-1*	*-1*	***+1***	*-1*	*-1*	*-1*

Indeed, the pathway is not predicted by the projection procedure. The first catabolic step of this pathway is catalyzed by the L-carnitine dehydrogenase. In order to search for this enzymatic activity in Pse*udomonas putida* KT2440, we used the "Search by keyword" tool in Microscope, which allows to perform searches in the functional annotations of the genes together with their various analysis results. In this case, we searched for *Pseudomonas putida* KT2440 proteins having significant BLASTP similarity with SwissProt-TrEMBL entries annotated as «L-carnitine dehydrogenase»; these entries could possibly linked to an experimental validation (EXP means that the corresponding UniProt entries are associated to a pubmedID) :

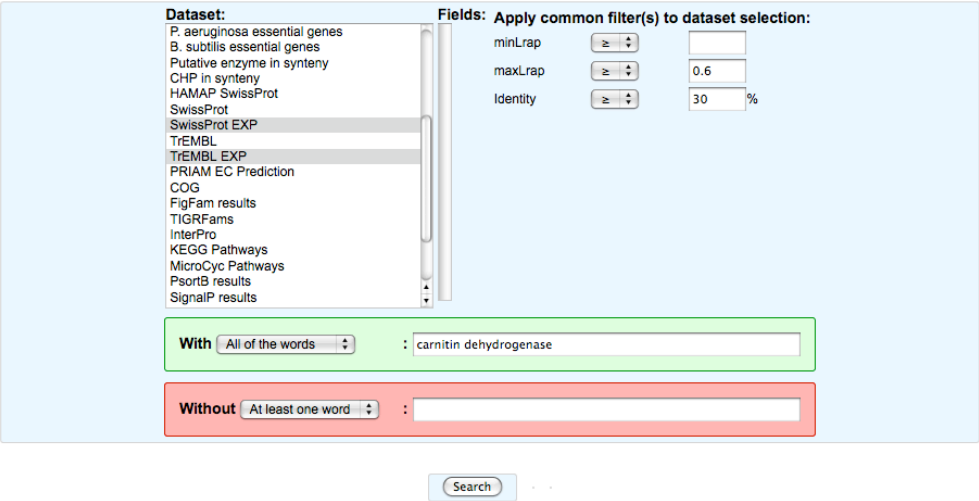

Figure 7: Snapshot of the Keyword Search tool of the Microscope platform – following similarity threshold has been used: at least 30% identity in amino acid between the two proteins; alignment on at least 60% of the length of the shortest protein.

The result of this search yields to 2 significant TrEMBL-EXP genes having significant sequence similarity with one *Pseudomonas putida* KT2440 genes: **PP0382**.

TrEMBL EXP [2]

Copy CSV Print

MoveTo	Label	PB id	Exp	maxLrap	minLrap	Ident %	Eval	OrderQ	OrderB	Gene	Description	EC number	Keywords	PubMedId
	PP0302	D7URM0	IPMed?	1	1	89.72	5.38273e-165	6	1	Xt-cdh	L-carnitine dehydrogenase	–	–	20530902
	PP0302	D7UNT2	IPMed?	0.637827	0.987539	48.9	4.51714e-79	12	1	lcdh	L-carnitine dehydrogenase	1.1.1.108	Oxidoreductase	20530902

Figure 8: Significant BLASTP results of *Pseudomonas putida* KT2440 genes (Label column) with TrEMBL entries annotated as L-carnitine dehydrogenase (Description column) with possibly experimental evidence in bibliography (PubMedId column)

Manual inspection of the BLASTP alignments and the significance of the similarity results (Eval column in Figure 5) indicates that PP0302, which was originally annotated as a '3-hydroxyacyl dehydrogenase' based on protein domain structure, is indeed the missing L-carnitine dehydrogenase enzymatic function. As a consequence, the annotation of this gene was edited accordingly in the Microscope environment, specifying the MetaCyc reaction associated with this enzymatic activity:

Text Format Help

Genomic Object Editor: PP0302

***Pseudomonas putida* KT2440 - chromosome PP NC_002947**

5'/3' TrEMBL alignments SwissProt alignments PhyloProfile PubMed KEGG BRENDA MicroCyc

CURRENT ANNOTATION MaGe curated annotation | Status: finished | Annotator: **adanchin**

Type	Begin	End	Length	Frame	Mutation	Gene	Synonyms	Date	Status
CDS	362515	363480	966 (321aa)	-1	no	cdh		2011-06-27 14:40:48	finished

Note	similar to 3-hydroxyacyl-CoA dehydrogenase family protein from Pseudomonas putida (strain KT2440), swall : Q88R32 (321 aa), Evalue = 1.07956e-181, %identity = 100, on 321 aa ; and similar to 3-hydroxyacyl-CoA dehydrogenase, NAD-binding from Pseudomonas putida (strain F1 / ATCC 700007), swall : A5VX85 (321 aa), Evalue = 3.47284e-180, %identity = 99.38, on 321 aa ; and similar to 3-hydroxyacyl-CoA dehydrogenase NAD-binding from Pseudomonas putida (strain W619), swall : B1JEE6 (321 aa), Evalue = 2.1069e-177, %identity = 96.26, on 321 aa.
Product	L-carnitine dehydrogenase
Comments	/Experimentally based evidence: activity demonstrated in Pseudomonas putida IFP 206, Xanthomonas translucens and Rhizobium sp.; note that the KM is approximately 1 mM
EC number	1.1.1.108
MetaCyc Reaction	CARNITINE-3-DEHYDROGENASE-RXN: carnitine 3-dehydrogenase
PubMedId	20530902, 3058208
Product Type	e : enzyme
Localization	1 : Unknown
Class	1b : Function experimentally demonstrated in the studied species
BioProcess	6.2 : Amino acids and amines
Roles	1.1.4.2 : Carnitine degradation

CANCEL SAVE

Figure 9: Final annotation of the PP0302 entry in Microscope platform based on the results of the manual curation. Specific Metacyc reaction associated to this enzymatic activity is reported.

The L-carnitine dehydrogenase converts L-carnitine to 3-dehydrocarnitine. The later part of the pathway, which includes the break down of 3-dehydrocarnitine to glycine betaine, and potentially to glycine, is not as well characterized: the reactions have not been characterized, and the enzymes responsible for them have not been identified. (Figure 10).

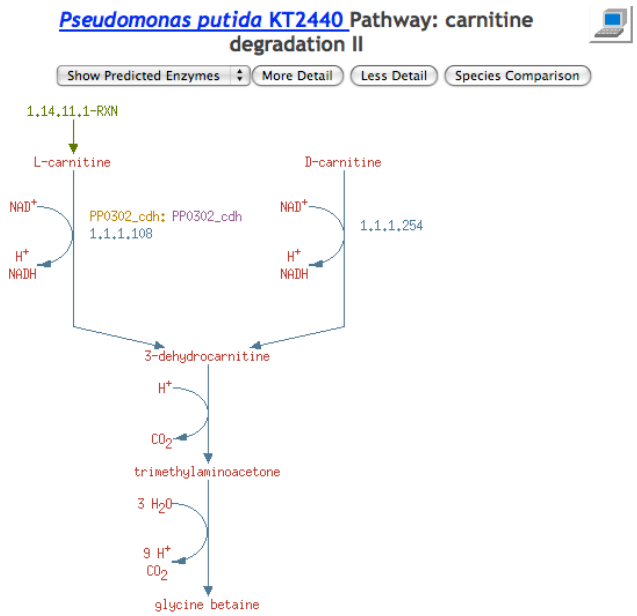

Figure 10: Carnitine degradation pathway variant II in Metacyc.

Pathway degradation are known to contain enzymatic activities encoded by genes found in the same transcription unit or at least, co-localized on the chromosome. Using both the synteny result and other MicroScope functionalities it was easy to explore the PP0302 gene context in order to search for candidate genes for these orphan enzymatic activities. For example, the co-evolution scores calculation allows to compare the phylogenetic profiles of the protein coding genes in a given pathway (patterns of presence/absence in other bacterial genomes) taking into account the phylogenetic distance and the gene order conservation between bacterial genomes (Engelen *et al*., submitted). Asking for "co-evolved genes for carnitine degradation pathway" in *Pseudomonas putida* KT2440 we found that the **PP0303** gene, annotated as "conserved protein of unknown function", is one of the best candidate (rank 1). In addition, the **PP0301** gene annotated as a "putative thioesterase" is another interesting gene candidate (rank 7).

Reactions in « *carnitine degradation II* »

Copy CSV Print

Reactions	EC Number(s)	Acinetobacter baylyi ADP1	Pseudomonas aeruginosa PAO1	Pseudomonas putida KT2440
1.1.1.254-RXN: (*S*)-carnitine 3-dehydrogenase	1.1.1.254	_	_	_
CARNITINE-3-DEHYDROGENASE-RXN: carnitine 3-dehydrogenase	1.1.1.108	_	_	PP0302
RXN-5882: RXN-5882	_	_	_	_
RXN-5881: RXN-5881	_	_	_	_

"carnitine degradation II" MicroCyc Cross-species comparison

Co-evolved genes for «*carnitine degradation II*»

Pseudomonas putida KT2440 [500]

Showing 1 to 10 of 500 results Show 10 Results Search: Copy CSV Print

Label	Gene	Coevolution score average	Rank average	Product
PP0303	_	0.837	1	conserved protein of unknown function
PP0323	soxB	0.801	2	Sarcosine oxidase subunit beta
PP5182	_	0.739	3	Aminotransferase, class III
PP0384	_	0.733	4	Transcriptional regulator, AsnC family
PP3572	_	0.732	5	conserved protein of unknown function
PP2836	_	0.728	6	Fumarylacetoacetate hydrolase family protein
PP0301	_	0.714	7	putative thioesterase

Figure 11: *Pseudomonas putida* KT2440 protein coding genes sorted by it´s co-evolution score average with PP0302 gene encoding L-carnitine dehydrogenase in the carnitine degradation pathway.

Indeed the synteny map (Figure 12) shows that the gene cluster containing PP0301 to PP0305 CDSs is absent in *E. coli*, *A. baylyi* and *S. thyphimurium*. It is present in *P. aeruginosa*, Burkholderia species and *Lutiella nitroferrum* (plus other different bacteria genomes). Such results show that these genes are functionally linked, and that they are very probably involved in the carnitine degradation.

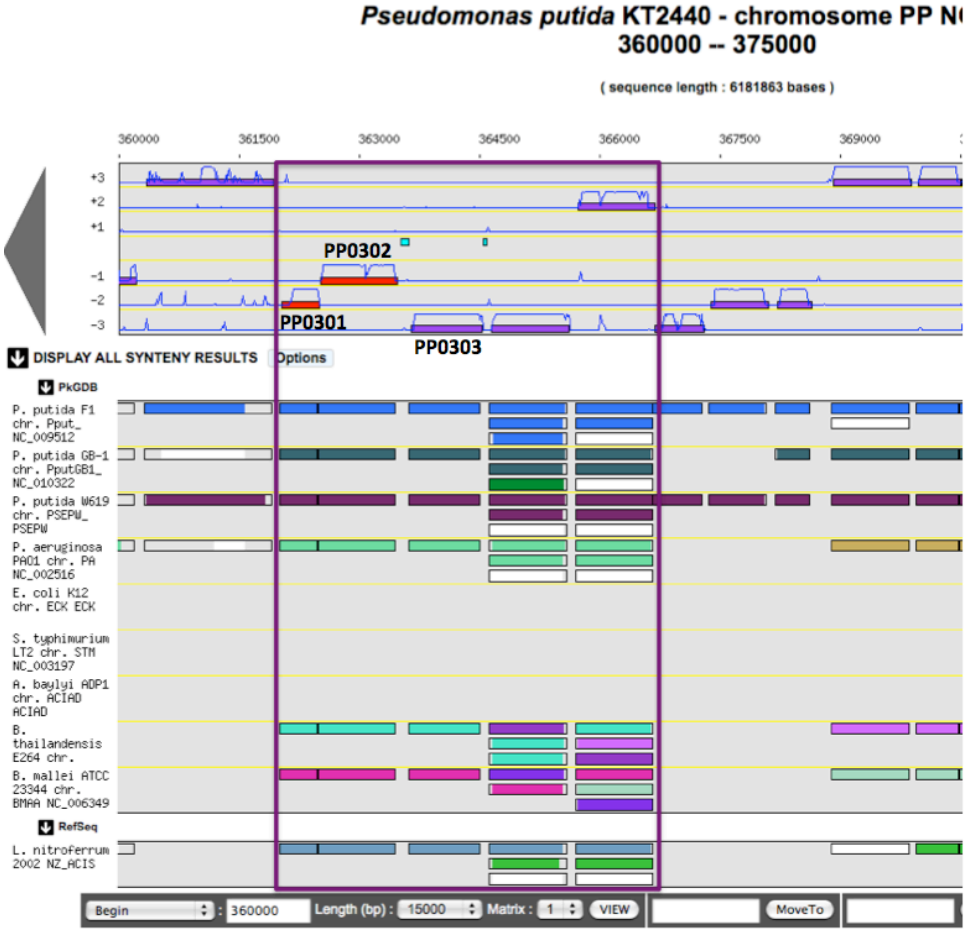

Figure 12: Synteny maps reveals that the PP0301-PP0305 gene cluster is conserved in several bacterial genomes

Experimental studies carried out at Genoscope (CEA) on this gene cluster provide novel evidences about the enzymes and reaction mechanisms involved in subsequent steps of L-carnitine degradation in *Pseudomonas putida* KT2440. These novel reactions and the Gene-Protein-Reaction assignments will be added into the Microscope platform as soon as the results are published. These new data will be available in the Microme database at the same time.

Perspectives

This first jamboree has been essential for attendees to understand what does “pathway curation” mean. Indeed, before entering and designing novel metabolic pathways in Microme, curator first need to search for gene candidates when reactions are missing and to find corresponding experimental validation of enzymatic activities if any. These curation steps involve other tools developed by Microme partners. Description of the Standard Operating Procedures needed to curate metabolic data (D2.4) will give a more precise picture of how these tools and these results could be use to ease this curation process.
Finally, this jamboree specifically aimed at preparing the Microme database for the first public release (D2.3).

References

1. Erin F Mulrooney et al., «Biosynthesis of UDP-N-acetyl-L-fucosamine, a precursor to the biosynthesis of lipopolysaccharide in *Pseudomonas aeruginosa* serotype O11 », The Journal of Biological Chemistry 280, nº. 20 (mai 20, 2005): 19535-19542.
2. D Vallenet et al., « MicroScope: a platform for microbial genome annotation and comparative genomics », Database: The Journal of Biological Databases and Curation 2009 (2009): bap021
3. Microscope : https://www.genoscope.cns.fr/agc/microscope
4. Microme Dev : http://dev.microme.eu
5. Engelen Stefan, Vallenet David, Médigue Claudine and Danchin Antoine « Distinct co-evolution patterns of genes associated to DNA polymerase III DnaE and PolC". Submitted (*BMC Genomics*)