Data integration with Life Science databases: the technical aspects

Alexander Vasilenko, Svetlana Ozerskaya, Oleg Stupar, Lujdmila Evtushenko ,
Paolo Romano, Linhuan Wu, Juncai Ma

Potential help for the Life Science world from microbial culture collections (mBRCs)

1. Assurance of repeatability of experimental data
2. Resolving nomenclatural issues related to microorganisms
3. Strain-specific characters

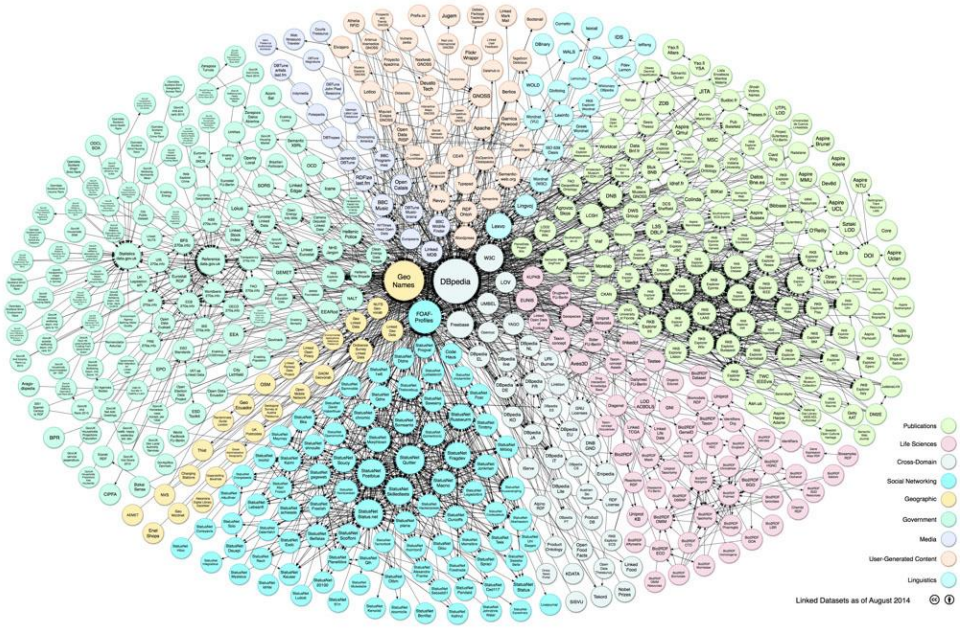
Microbial databases – CC interconnection

Task 1

Task2

GCM/ WDCM

1116 databases



111/725

<10%

0%

725 Catalogues

Database has 10 partners on average

Catalogue has 0 partners in average

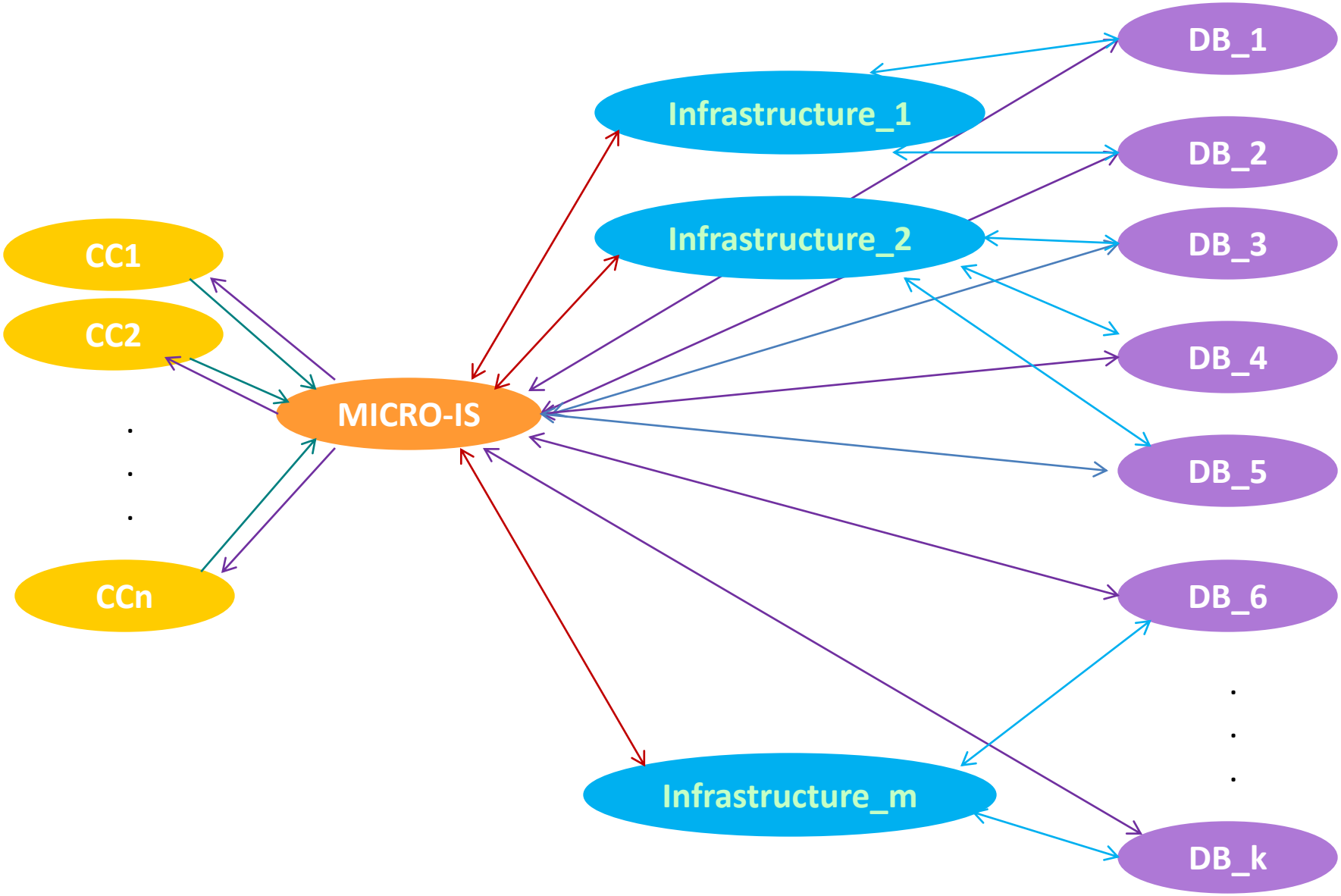
Potentially this data integration could mean the tasks:

1. To make CC data visible and accessible from partner Life Science databases,
2. To make partner database records visible and accessible from CC aggregated catalogue,

in the formats:

a. To give this data integration for human access,
b. To give this data integration for computer programs.

General integration schema



Life Science databases inspected

Total number of life science database names or references discovered in this study is more than 14 800. The total number of database references inspected manually is more than 5500 . The total number of life science databases collected visible online is 2660, the number of databases with microbial data collected is 1116 (plus 9318 bacterial databases in BioCyc).

The main sources:

- MB (1802 entries (http://metadatabase.org), 26.12.2015),
- Biosharing (724 databases, (https://www.biosharing.org/), 26.12.2015),
- BioMedBrigeds (814 Databases, 27.12.2015), (http://wwwdev.ebi.ac.uk/fgpt/toolsui/),
- Pathguide (363 database names, 2013), (http://www.pathguide.org/)
- ELIXIR list (579 entries, (https://bio.tools/?q=database), 28.1.2016)
- ExPASy (85 + 665 databases, (http://www.expasy.org/old_links), 12.2.2016)
- Bioinformatics Links Directory (621 databases)
- OBRC (http://www.hsls.pitt.edu/obrc/) 30.3.2017

Databases parameters collected (an example)

- Unique identifier: BIODBCORE-000438
- Database acronym: dbSNP
- Database name: The Database of Short Genetic Variation (single nucleotide polymorphism)
- Database URL: http://www.ncbi.nlm.nih.gov/SNP/
- Access level: Open
- Practical domain: health
- Microbial level: sp
- Year of the last correction: 2017
- Developer/Owner: USA, NCBI; USA, National Library of Medicine, National Institutes of Health
- Comment: Escherichia coli
- Orientation: -
- Properties: DNA, gene, genome, proteomics, publications, RNA
- Search by: -
- Ontologies list: SO
- Partner databases: Assembly, BioProject, BioSample, ClinVar, dbGaP, dbMHC, dbSTS, dbVar, Ensembl, GenBank, Homologene, IGSR, MapViewer, NCBI Gene, Nucleotide, OMIM, PMC, Protein, PubChem Substance, PubMed, RDP, RefSeq, UniGene, UniProtKB
- Program interface: ELIXIR WEB UI, Entrez Programming Utilities (E-Utils)

Property keywords in databases

826	Gene	226	Interactome
716	Proteomics	219	Taxonomy
625	Publications	195	Drugs
517	Image	195	Peptide
389	RNA	183	Molecules
395	DNA	166	Metabolite
355	Genome	151	Pathogen
361	Enzyme	167	Immunology
316	Cell	156	Toxicology
297	Chemistry	132	Lipid
270	Pathways	45	Microbiome
263	Disease		

Average number of keywords assigned to a database = 6,55

Databases with microbiome data

Biology Reference
BioSamples
BioSystems
Bookshelf
EMBL
EMBL-EBI
ENA
Espacenet
Europe PMC
ForestScience Current Database
GO Database
GONUTS
GoPubMed
HGTree
HOMD
HPMCD
IMG
IMG/M
IMG/VR
MEDLINE
MeSH
Microbiome
NARCIS
NCBI
NFSD
NLM Catalog
Nowomics
OMIM
OMIM (1)
OReFiL
PANGAEA
PLOS One
PMC
PSP
pubget
PubMed
PubMed Health
QIAGEN
RefSeq
ScienceDirect
SRA
TACONIC
UniProtKB
VetMed Resource
WikiGenes

Biggest database producer: BESC (BioEnergy Science Center)

BioCyc pathway/genome database: 9367 databases totally (http://www.biocyc.org/biocyc-pgdb-list.shtml)

Group 1 are 7 databases: **EcoCyc**, **MetaCyc**, **HumanCyc**, **AraCyc**, **YeastCyc**, **LeishCyc**, **TrypanoCyc**

Group 2 are 41 databases generated by program with curation done each is one strain:

- Agrobacterium fabrum C58
- Anopheles gambiae
- Aurantimonas manganoxydans SI85-9A1
- Bacillus anthracis Ames
- Bacillus subtilis 168
- Bacteroides thetaiotaomicron VPI-5482
- Candidatus Cardinium hertigii
- Candidatus Evansia muelleri
- Candidatus Portiera aleyrodidarum BT-QVLC
- Caulobacter crescentus CB15
- Caulobacter crescentus NA1000
- Chlamydomonas reinhardtii
- Clostridium saccharoperbutylacetonicum ATCC 27021
- Cryptosporidium hominis TU502
- Cryptosporidium parvum Iowa
- Drosophila melanogaster
- Escherichia coli B str REL606
- Escherichia coli CFT073
- Escherichia coli K-12 substr W3110
- Escherichia coli O157:H7 str EDL933
- Eubacterium rectale ATCC 33656
- Helicobacter pylori 26695
- Listeria monocytogenes 10403S
- Methylosinus trichosporium OB3b
- Mus musculus
- Mycobacterium tuberculosis CDC1551
- Mycobacterium tuberculosis H37Rv
- Penicillium chrysogenum Wisconsin 54-1255
- Peptoclostridium difficile 630
- Plasmodium berghei ANKA
- Plasmodium chabaudi
- Plasmodium falciparum 3D7
- Plasmodium vivax Sal-1
- Plasmodium yoelii 17XNL
- Schistosoma mansoni
- Shigella flexneri 2a str 2457T
- Streptomyces coelicolor A3(2)
- Synechococcus elongatus PCC 7942
- Thalassiosira pseudonana CCMP1335
- Toxoplasma gondii ME49
- Vibrio cholerae O1 biovar El Tor str N16961

Group 3 are 9318 databases each database is one bacterial strain with no curation yet

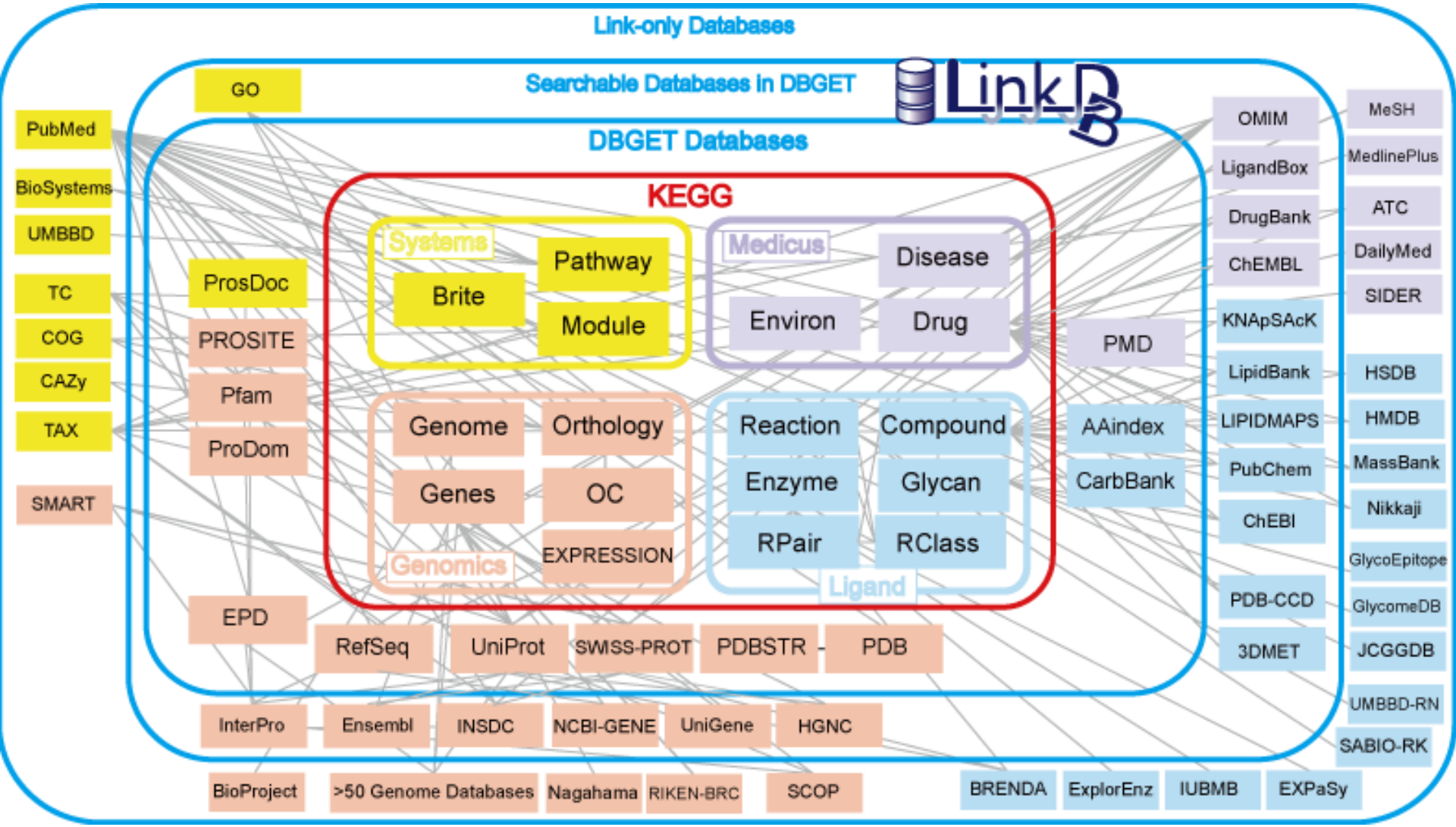
EMBL-EBI (97 dbs)

ArrayExpress
ASD
ASTD
ATD
BioModels
BioSamples
Cellular Phenotype Db
ChEBI
ChEMBL
CluSTr
CSA
DGVa
DNAtraffic
DrugPort
e!Ensembl
e!Ensembl S. cerevisiae
e!EnsemblBacteria
e!EnsemblCat
e!EnsemblChicken
e!EnsemblChimpanzee
e!EnsemblCow
e!EnsemblDog
e!EnsemblFugu
e!EnsemblFungi
e!EnsemblGenomes
e!EnsemblGorilla
e!EnsemblHorse
e!EnsemblMetazoa
e!EnsemblMouse
e!EnsemblPig
e!EnsemblPlants
e!EnsemblProtists
e!EnsemblRabbit
e!EnsemblZebrafish
EGA
EMBL
EMBL-EBI
EMDB
ENA
Ensembl
Enzyme Portal
Enzyme Structures
EVA
Expression Atlas
FunTree
GeneDB
GWAS Catalog
HGNC
HipSci
IGSR
IMEx
IMGT/HLA
IntAct
IntEnz
InterPro
IPD
IPD-ESTDAB
IPD-HPA
IPD-KIR
IPD-MHC
logRECOORD
MACiE
MEROPS
MetaboLights
Metal MACiE
MicroCosm
MIRIAM collection
MTBLS
NRNL1
NRNL2
NRPL1
NRPL2
OLDERADO
PANDIT
PDBe
PDBe EM Resources
PDBeChem
PDBsum
Pfam
Pfam
PhenoDigm
PICR
PomBase
PRIDE
PROCOGNATE
Reactome
RECOORD
Rfam
RNAcentral
SAS
SRS@EMBL-EBI
SureChEMBL
TreeFam
UniChem
UniProt-GOA
UniSave
VASCO
VectorBase

NCBI (70 dbs)

Assembly
BioProject
BioSample
BioSystems
Bookshelf
CCDS
CDD
ClinGen
ClinVar
Clone DB
COGs
dbEST
dbGaP
dbMHC
dbProbe
dbSNP
dbSTS
dbVar
Dengue virus database
ECRbase
Genbank
Gene
Genetic Codes
Genome
GEO
GEO DataSets
GEO Profiles
GSS
GTR
Histone
HIV-1
Homologene
IBIS
Influenza Virus Resource
MapViewer
MedGen
MEDLINE
MeSH
MMDB
NCBI
NCBI taxonomy
NCBI Trace Archives
NLM Catalog
Nucleotide
OMIM
Organelle genomes
PMC
PopSet
Probe
Protein
Protein Clusters
PubChem
PubChem BioAssay
PubChem Compound
PubChem Substance
PubMed
PubMed Health
RefSeq
RefSeqGene
Retroviruses
SKY/M-FISH and CGH
SPARCLE
SpliceInfo
SRA
Structure
TPA
UniGene
UniVec
Viral genomes
Virus Variation

Databases interconnection example



Interconnection matrix

	3D Lectin	3D RIBOSC	5S RNA Da	A.pernix	UniPROBE	UniProt-G	UniProtKB	UniRef	YMPL	YPM	Zif-BASE	AN
2P2Idb	0	0	0	0	0	0	0	0	0	0	0	1
3D Genome Browser	0	0	0	0	0	0	0	0	0	0	0	1
3D Lectin	0	0	0	0	0	0	0	0	0	0	0	0
3D RIBOSOMAL MODIFICATIO	0	0	0	0	0	0	0	0	0	0	0	3
3DBIONOTES	0	0	0	0	0	0	0	0	0	0	0	0
3Dee	0	0	0	0	0	0	0	0	0	0	0	0
3DEM Benchmark	0	0	0	0	0	0	0	0	0	0	0	0
3D-footprint	0	0	0	0	0	0	0	0	0	0	0	1
3DID	0	0	0	0	0	0	0	0	0	0	0	6
PubChem	0	0	0	0	0	0	0	0	0	0	0	57
PubChem BioAssay	0	0	0	0	0	0	0	0	0	0	0	16
PubChem Compound	0	0	0	0	0	0	0	0	0	0	0	31
PubChem Substance	0	0	0	0	0	0	0	0	0	0	0	25
pubget	0	0	0	0	0	0	0	0	0	0	0	2
Public Data Portal	0	0	0	0	0	0	0	0	0	0	0	1
PubMed	0	0	0	1	0	0	1	0	0	0	1	565
PubMed Health	0	0	0	0	0	0	0	0	0	0	0	6
YPM	0	0	0	0	0	0	0	0	0	0	0	1
ZFIN	0	0	0	0	0	1	1	0	0	0	0	20
Zif-BASE	0	0	0	0	0	0	0	0	0	0	0	0
ZOBODAT Vespoidea	0	0	0	0	0	0	0	0	0	0	0	1
ZorapteraSF	0	0	0	0	0	0	0	0	0	0	0	1
ИИПС	0	0	0	0	0	0	0	0	0	0	0	0
CN	5	8	4	10	6	148	161	15	2	3	5	

UniProtKB database has integration parameters CN=161, AN=501.

In CC catalogues both values are mostly 0.

Databases with more than 26 partners (CN)

364 Pathguide
161 UniProtKB
159 iProClass
153 COL
148 UniProt-GOA
143 OReFiL
129 EcoliWiki
128 GeneCards
91 PIR
73 UCD 2D-PAGE
63 Hits
61 SWISS-2DPAGE
57 PubChem
57 PubChem BioAssay
57 PubChem Compound
57 PubChem Substance
55 dbProbe
55 NCBI
53 PiroplasmsDB
52 e!EnsemblGenomes
49 EMBL
49 ENA
49 SBKB
46 EcoGene
46 SGD
45 Gene
44 InterMitoBase
43 EMBL-EBI
43 OMIM
43 OMIM (1)
42 MetaCyc
40 Guide to Pharmacology
40 MalaCards
38 NCBI taxonomy
38 OpenHelix
38 ViralZone
37 GPCRs
37 T3DB
36 EcoCyc
36 Nucleotide
35 BacMet
35 ChEBI
35 PRODORIC
34 ESTHER
33 Genome
33 gpmDB
32 Ebolavirus
32 HOGENOM
32 STITCH
31 eNet
31 HPIDB
31 PANDORA
30 FungiDB
30 JPGV
30 Reactome
30 STRING
29 BioModels
29 CAZy
29 ConsensusPathDB
29 EuPathDB
29 MACiE
29 RefSeq
28 EcoProDB
28 RNAcentral
28 ThaleMine
28 YeastMine
27 BioSystems
27 OrthoDB

Databases with big attraction number (AN)

565 PubMed
501 UniProtKB
275 NCBI taxonomy
255 RCSB PDB
239 Genbank
229 Gene
199 KEGG
193 RefSeq
187 EC
182 Pfam
160 InterPro
157 Protein
152 Ensembl
125 Nucleotide
109 OMIM
99 SGD
80 ENA
80 HGNC

76 PROSITE
71 CAS
71 IntAct
70 Reactome
69 ChEBI
63 FlyBase
61 MEDLINE
61 UniGene
59 BioGRID
57 PubChem
56 GEO
56 NCBI
56 SMART
55 DrugBank
55 MGI
55 PIR
55 PMC
55 SCOP

54 COGs
53 Genome
51 DIP
51 STRING
50 BioCyc
50 ENZYME
49 KEGG PATHWAY
48 HPRD
48 WormBase
47 BioProject
47 BRENDA
47 DDBJ
47 TAIR
47 TIGRFAMS
46 MINT
45 MeSH
43 PANTHER
40 GeneCards

38 CATH
38 MetaCyc
36 CDD
36 SUPERFAMILY
35 dbSNP
35 EcoCyc
34 ChEMBL
34 PDBsum
33 PLOS One
33 PRINTS
31 Homologene
31 ProDom
31 PubChem Compound
31 wwPDB
30 PDBe
30 UCSC Archaeal Genome Browser

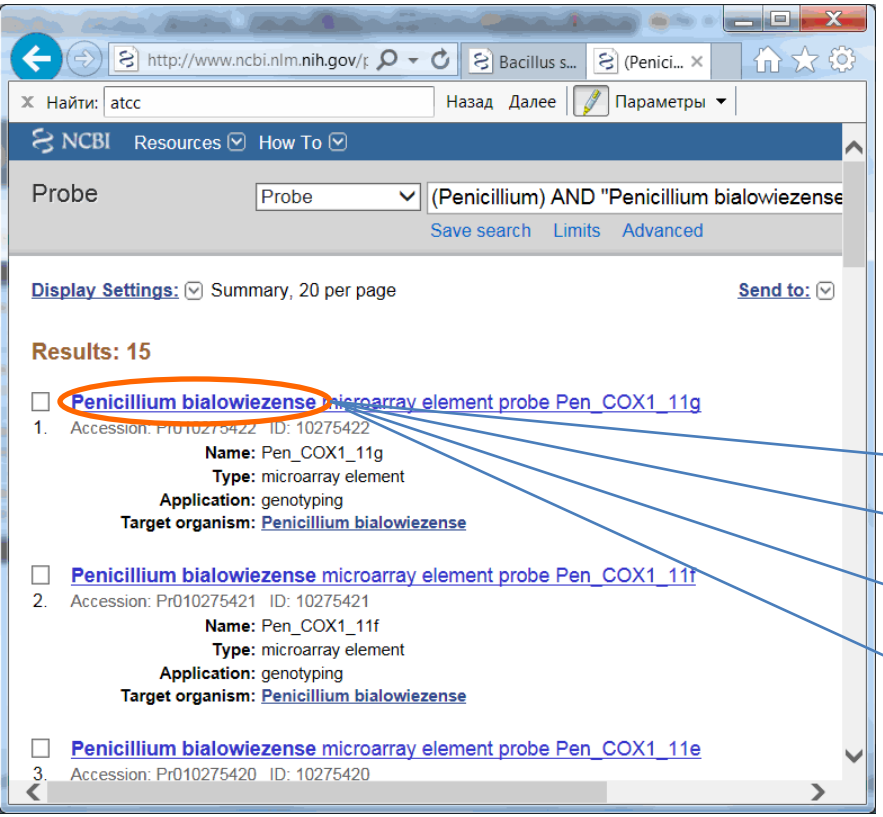
Integration candidates

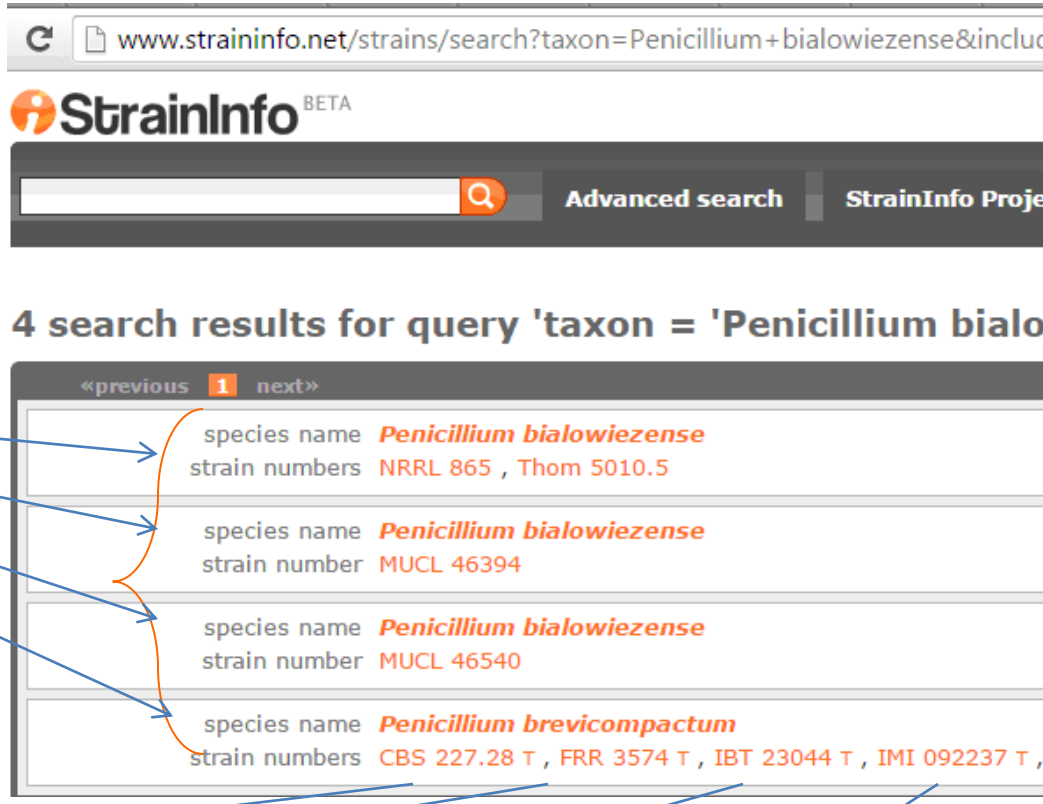
	Producer	Databases	Total attraction	% of maximal
1	NCBI	70	2909	33
2	EMBL-EBI	97	1209	14
3	SIB	37	762	9
4	Kioto University	19	348	4
5	Instute Paster	18	148	2
6	BioCyc	9378	133	2
7	InterMine	16	20	0
	1+2+3+4	133	5228	59
	1+3+4	92	4019	45
	2+3+4	78	2319	26
	Total	1115	8870	100

MICRO-IS database partners in efficient solution

ArrayExpress, Assembly, BioModels, BioProject, BioSample, BioSamples, BioSystems, Bookshelf, CDD, Cellosaurus, ChEBI, ChEMBL, COGs, CSA, dbEST, dbProbe, dbSNP, Dengue virus database, DNAtraffic, DrugPort, e!Ensembl, e!Ensembl Saccharomyces cerevisiae, e!EnsemblBacteria, e!EnsemblFungi, e!EnsemblGenomes, e!EnsemblProtists, EMBL, EMBL-EBI, EMDB, ENA, Ensembl, ENZYME, Enzyme Structures, EPD, EVA, Expression Atlas, Genbank, Gene, GeneDB, Genetic Codes, Genome, GEO, GEO DataSets, GEO Profiles, GSS, HAMAP, Hits, HIV-1, Homologene, IMEx, Influenza Virus Resource, IntAct, InterPro, KEGG, KEGG BRITE, KEGG DISEASE, KEGG GENES, KEGG GENOME, KEGG GLYCAN, KEGG LIGAND, KEGG MEDICUS, KEGG MODULE, KEGG Organisms, KEGG ORTHOLOGY, KEGG PATHWAY, MACiE, MapViewer, MedGen, MEDLINE, MEROPS, MeSH, MetaboLights, MIAPEGelDB, MMDB, MTBLS, NCBI, NCBI taxonomy, NCBI Trace Archives, neXtProt, NLM Catalog, Nucleotide, OMA, OMIM, OpenFlu, Organelle genomes, PathComp, PathPred, PathSearch, PaxDB, PDBe, PDBe EM Resources, PDBsum, Pfam, PICR, PMC, PMP, PomBase, PopSet, PRIDE, Probe, PROSITE, Protein, Protein Clusters, Protein Spotlight, Proteomes, PubChem, PubChem BioAssay, PubChem Compound, PubChem Substance, PubMed, PubMed Health, Reactome, RefSeq, RefSeqGene, Retroviruses, Rfam, Rhea, RNAcentral, SPARCLE, SpliceInfo, SRA, Structure, SugarBind, SWISS-2DPAGE, SWISS-MODEL, SwissVar, UniGene, UniProt-GOA, UniProtKB, UniRef, Viral genomes, ViralZone, Virus Variation

Task 1a:
Strains algorithm



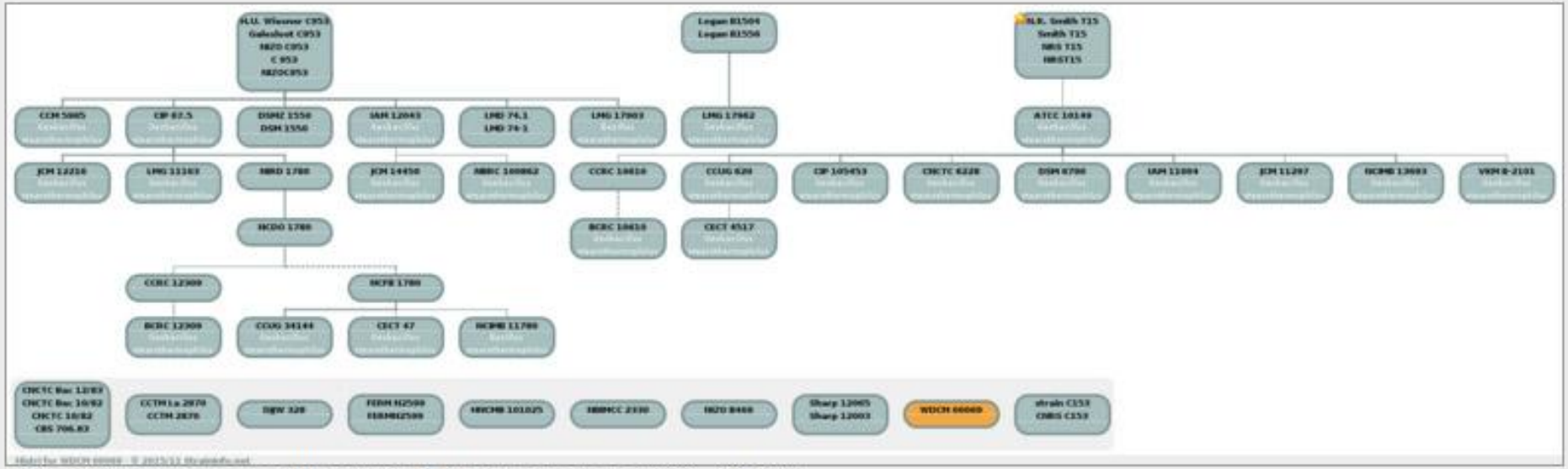


WDCM 133 Centraalbureau voor Schimmelcultures Filamentous fungi and Yeast Collection, Netherlands	WDCM 18, Food, Science, Australia, Ryde	WDCM 758, IBT, Culture Collection of Fungi, Denmark	WDCM 214, CABI, Genetic Resource Collection, UK

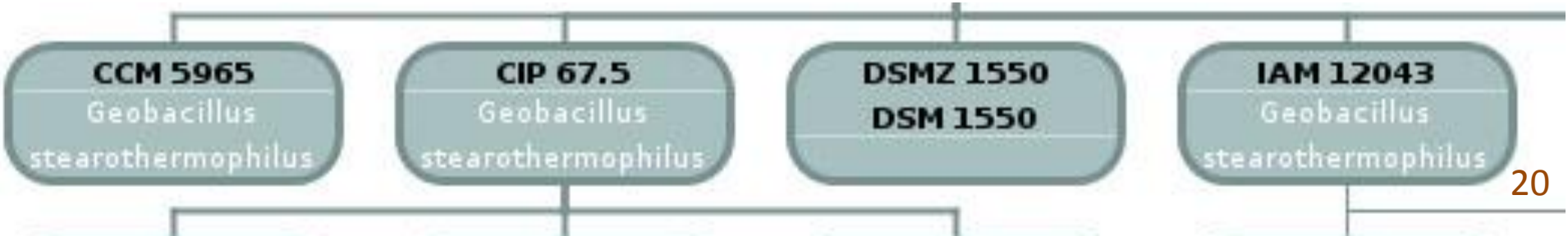
StrainInfo strain exchange tree

Strain Passport
WDCM 00069 *Geobacillus stearothermophilus*

history



This Histri was manually curated (last saved on 2011-04-12).



Solutions: task 1a: Name processing *

Penicillium cyaneofulvum

Summary:	Penicillium cyaneofulvum Biourge, La Cellule 33: 174 (1923)
Synonymy:	=Penicillium brunneorubrum Dierckx, Annales de la Société Scientifique de Bruxelles 25 (1): 88 (1901) =Penicillium griseoroseum Dierckx, Annales de la Société Scientifique de Bruxelles 25 (1): 86 (1901) =Penicillium chrysogenum Thom, U.S.D.A. Bureau of Animal Industry Bulletin 118: 58 (1910) =Penicillium baculatum Westling, Svensk Botanisk Tidskrift 4: 139 (1910) =Penicillium notatum Westling, Arkiv før Botanik 11 (1): 95 (1911) . . . =Penicillium fluorescens Laxa, Zentralblatt für Bakteriologie und Parasitenkunde Abteilung 2 86 (5-7):160-165 (1932) =Penicillium camerunense R. Heim, Bull. Acad. R. Belg. Cl. Sci.: 42 (1949) =Penicillium aromaticum f. microsporum Romankova, Uchenn. Zap. Leningr. Univ. Zhadanov: 102 (1955) =Penicillium harmonense Baghd., Novosti Sistematiki Nizshikh Rastenii 5: 102 (1968)
Current name:	Penicillium chrysogenum Thom, U.S.D.A. Bureau of Animal Industry Bulletin 118: 58 (1910)
Classification:	Fungi, Ascomycota, Pezizomycotina, Eurotiomycetes, Eurotiomycetidae, Eurotiales, Trichocomaceae, Penicillium
Facultative or heterotypic synonyms:	1. Penicillium aromaticum f. microsporum Romankova, Uchenn. Zap. Leningr. Univ. Zhadanov: 102 (1955) 2. Penicillium baculatum Westling, Svensk Botanisk Tidskrift 4: 139 (1910) 3. Penicillium brunneorubrum Dierckx, Annales de la Société Scientifique de Bruxelles 25 (1): 88 (1901) 4. Penicillium camerunense R. Heim, Bull. Acad. R. Belg. Cl. Sci.: 42 (1949) 5. Penicillium chlorophaeum Biourge, La Cellule 33: 271 (1923) 6. Penicillium chrysogenum Thom, U.S.D.A. Bureau of Animal Industry Bulletin 118: 58 (1910) . . .

* Page content from: http://www.mycobank.org/BioloMICS.aspx?Table=Mycobank&Rec=18759&Fields=All

TAXONOMY IN MICROBIAL DATABASES

Taxonomy	*	**
NCBI	115	267
GBIF	16	16
IF	7	9
COL	6	6
LPSN	2	2
MycoBank	2	5

* References in databases with taxonomical data

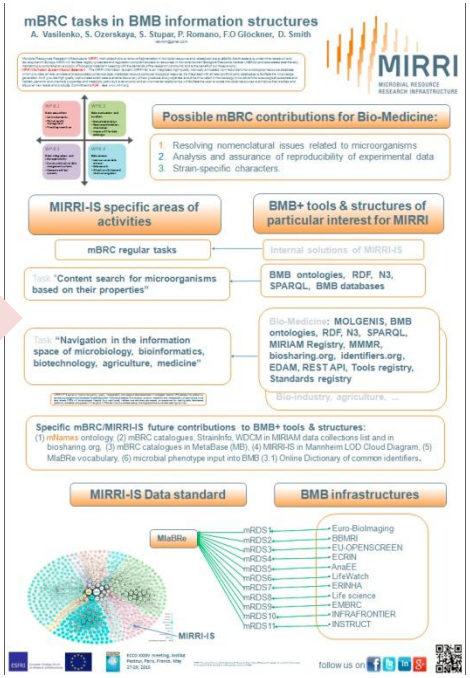
** In all the databases with microbial data

229 - Total number of databases with taxonomical data

94 - No reference to specific taxonomy database

Task 2: solution tools

Producer	Interface
NCBI	Entrez Programming Utilities (E-Utils)
EMBL-EBI	RESTful Web Services interface, Semantic WEB, RDF, SPARQL endpoint
SIB	RESTful Web Services interface, Semantic WEB
Kioto University	KEGG API, LinkDB



Thank you