

StrainInfo.net: breaking down information barriers into holistic data integration scenarios using globally unique identifiers

Peter Dawyndt^{a,b,*}, Bernard De Baets^b, Xianhua Zhou^c, Juncai Ma^c, Jean Swings^{a,d}

^a Laboratory of Microbiology, Ghent University, Belgium
^b Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Belgium
^c Information Network Center, Institute of Microbiology, Chinese Academy of Sciences, PR China
^d BCCMTM/LMG Bacteria Collection, Ghent University, Belgium
 *corresponding author: Peter.Dawyndt@UGent.be

中国科学院微生物研究所
 Institute of Microbiology, Chinese Academy of Sciences

Mission statement

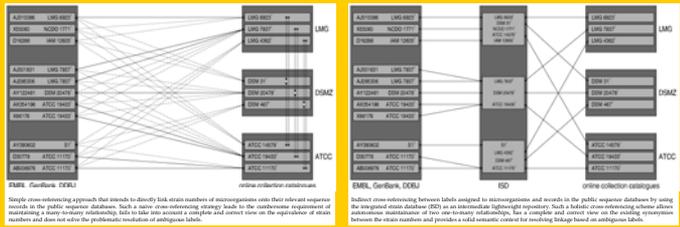
Once a microbial strain gets isolated from its natural habitat, a wealth of information about its genome, proteome, metabolism, clinical and ecological traits can be collected, on which basis it might eventually turn out to become important reusable material for scientific and industrial purposes. Perhaps a culture of the biological material gets deposited into a biological resource center (BRC) for long-term preservation and global dissemination among other BRCs or research institutions, its raw observational data are stored into private or public repositories to establish large-scale identification applications, it becomes commonly accepted as key reference material to support some artificial (human-conceived) taxonomic framework designed as a higher-level conception of biological diversity, it forms a cornerstone for the implementation of an industrial process that is protected by the patent law, or some conclusions drawn from the knowledge attained during scientific research activities wherein the microbial strain was involved are bundled into scientific publications.

Both the actual content of this downstream information on the microorganism and its location in private databases or on the World Wide Web are sensitive to modification over time. As science and technology are moving rapidly, thereby increasingly making use of the scientific merits of previous research results, instant and effortless visibility of this creative and scientific downstream information has become imperative for the realization of successful innovation chains that take full opportunity of the exploitation of biological resources.

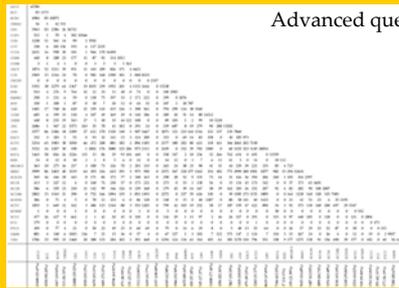
The StrainInfo.net portal (www.straininfo.net) therefore envisages the establishment of a technology platform that can stimulate this movement towards using multi-perspective integrated information in a broadened biological and clinical context, as indeed we all would rather like to benefit from automated ICT technologies for keeping track of downstream information on biological resources than putting all our efforts into the tedious and error-prone compilation of relevant knowledge from the heterogeneous and autonomous data collections spread across the information jungle.

What genes are sequenced of a given microorganism ?

- strain-level searches in public sequence databases not accurate enough
- simple integration approach not aware of existing synonyms between labels
- simple integration approach hampered by homonymous strain numbers
- holistic integration scheme of StrainInfo.net performs accurate parallel searches with all known strain numbers in the Integrated Strain Database
- incorporating references to the biological resources from the public sequence databases establishes a true divide and conquer strategy



Advanced queries (workflows) crossing database barriers



More accurate statistics based on integrated information

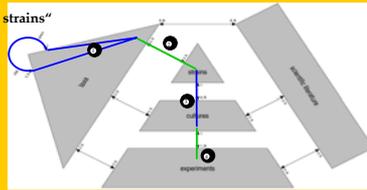
For a selection of biological resource centres (BRC), the table on the left gives a detailed overview of the number of isolates that each pair of BRCs has in common. Additionally the bottom of the table shows a reliable estimation of the number of strains that are unique within each collection^(*).

(*) acronyms of culture collections are taken from the World Data Centre of Microorganisms, taken into account the migration of culture collections and changes in culture collection acronyms.

Example query: "Find all 16S rRNA sequences of *Enterococcus* type strains"

4-step query resolution (right panel)

- 1) find all species of the genus *Enterococcus* (including synonyms)
- 2) find type strain for each species
- 3) find all synonym labels assigned to each type strain
- 4) find all 16S rRNA sequences linked to each of the synonym labels



Integrated Strain Database

Features & philosophy

- automatic assignment of globally unique identifiers to microbial strains
- maintenance of equivalence relation of strain numbers (machine learning)
- error detection/correction of inconsistencies within underlying sources
- provision of semantic context for disambiguation of label resolution
- dynamic integration platform for search results on microbial resources

Facts & Figures

- integrates information of 42 biological resource centres (culture collections)
 - organisms within scope: bacteria, archaea, filamentous fungi, yeasts
 - geographically distributed over all continents worldwide
 - ranging from niche specific to general purpose collections
- # organisms (isolates): 260,444
- # strain numbers (labels): 589,157
- # bacterial type strains affected by errors: 768/6137 (12.51%)

What collections do have a given microorganism in their holdings ?

information resource	country	WDCM	strain acronyms	records	homepage
American Type Culture Collection	U.S.A.	ATCC	ATCC, ATCC BAA, ATCC MYA	44,975	www.atcc.org
BIVTEC Culture Collection	Thailand	285	WCC	1,100	www.bivtec.org.th
Bioresource Collection and Research Center	Taiwan	89	BCRC	102,335	www.brc.gov.tw
Centre for Research and Biotechnology	Belgium	42	CBM	1,174	www.cbmr.be
Canadian Collection of Fungal Cultures	Canada	130	CCM	10,042	www.ccm.ca
Centre for Microorganisms and Cell Culture	Czech Republic	85	CCM, CCM F	3,427	www.mimik.cz
Microorganism Collection of the University of Microorganisms	Spain	80	CMIM	7,673	www.cimim.csic.es
Centre de Culture Typique	Brazil	32	CCCT	2,189	www.cct.org.br
Centre for Microorganisms and Cell Culture	Sweden	35	CMCC	16,928	www.ccmcc.se
Colectiunea Nationala de Culturi Microbiologice si Culturi Celulare	Mexico	500	CEBIM	16,025	www.ccmexico.mx
Centre for Microorganisms and Cell Culture	Spain	111	CCM	4,254	www.ccm.es
Collection de Cultures Typiques	France	789	CCP, CCP A	78,512	www.ccp.fr
Deutscher Sammlungszentrum für Mikroorganismen und Zellkulturen	Germany	274	DSMZ	14,628	www.dsmz.de
Food Science Australia, Inc.	Australia	14	FSM	2,949	www.foodscience.gov.au
TEAMBI Culture Collection	Finland	759	TEAMBI	2,089	www.teambi.com
International Collection of Micro-Organism Cultures and Cell Cultures	Japan	100	IAM	714	www.iam.ac.jp
IAM Culture Collection	Japan	100	IAM	3,120	www.iam.ac.jp
International Collection of Micro-Organism Cultures and Cell Cultures	Spain	149	ICM	14,925	www.icm.es
BCCM TM /JHIM - Biomedical Culture and Yeasts Collection	Belgium	442	JHIM	4,442	www.jhim.be
Japan Collection of Microorganisms	Japan	80	JCM	7,914	www.jcm.ac.jp
Korean Agricultural Culture Collections	South Korea	80	KACC	12,959	www.kacc.or.kr
Korean Collection for Type Cultures	South Korea	507	KCTC	7,015	www.kctc.or.kr
BCCM TM /LMG Bacteria Collection	Belgium	296	LMG	13,982	www.lmg.ac.be
BCCM TM /MCC - Mycological Culture Collection of Leuven	Belgium	308	MCC	11,993	www.mcc.ac.be
Microorganism Collection of Mexico	Mexico	85	CMC	1,113	www.cmc.mx
National Bank for Industrial Microorganisms and Cell Cultures	Belgium	138	NIMCC	466	www.nimcc.be
NITE Biological Resource Center	Japan	425	NIMC	12,959	www.nimc.ac.jp
National Collection of Agricultural and Industrial Microorganisms	Finland	485	NAM, N, NAM F, NAM Y	2,433	www.nam.ac.fi
National Collection of Industrial Microorganisms	India	3	NIM	2,002	www.nim.ac.in
National Collection of Industrial Food and Marine Bacteria	U.K.	63	NIMB	7,704	www.nimb.ac.uk
National Collection of Yeast Cultures	U.K.	481	NCYC	2,706	www.ncyc.ac.uk
Former Culture Collection of Cyprobiotechnia	France	481	CCP	452	www.ccp.fr
Persian Type Culture Collection	Iran	124	PTCC	322	www.ptcc.ac.ir
University of Alborz Microorganism Collection and Herbarium	Canada	78	UMC	452	www.umc.ca
All-Russian Collection of Microorganisms	Russian Federation	542	VIM, VIM A, VIM F, VIM Y	61,292	www.vim.ru
Common Access to Biological Resources and Information	Taiwan	1	CABRI	30,620	www.cabri.org