




Omnicrobe

Une base de données scientifique en libre accès sur les habitats et phénotypes microbiens

Sophie Schbath (MaIAGE)

Journée d'étude de l'assemblée des partenaires de Hal, le 21 novembre 2024



1


Un outil puissant pour les chercheurs en microbiologie

(<https://omnicrobe.migale.inrae.fr>)

- Accès à des données microbiennes diverses : centralisation de sources bibliographiques, collections microbiennes, bases génétiques
- Accélération et simplification de la recherche : facilite l'accès et le croisement d'informations normalisées pour des analyses avancées

Omnicrobe is a database of habitats, phenotypes and uses

The Omnicrobe database gathers comprehensive information on habitats, phenotypes and usages of microorganisms as automatically extracted from text sources (PubMed, GenBank, DSMZ, CIRM-BA, CIRM-CFBP, CIRM-Levures). The search offers you access to the information and their sources. [More information.](#)



- Accès : interface web et programme (API)
- Une action pluridisciplinaire : microbiologie, recherche et ingénierie en TAL, bioinformatique, système d'information

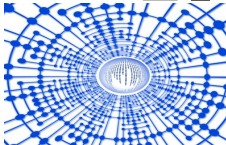
Omnicrobe
21 nov 2024/ AG Hal / S. Schbath

2

Un service d'accès à l'information scientifique en microbiologie


Normalisation des informations : extrait et normalise les informations relationnelles des textes et les relie à des référentiels

- Ontobiotope (habitats, phénotypes et usages)
- Taxonomie du NCBI pour les espèces et collections pour les souches



Gérer la diversité des données : analyse des sources décentralisées, semi- et non-structurées et les relie

- références bibliographiques de PubMed
- descriptions de souches microbiennes de collections (DSMZ, CIRM),
- description des milieux d'isolation d'espèces et souches (GenBank)



Omnicrobe
21 nov 2024/ AG Hal / S. Schbath

<https://agroportal.lirmm.fr/ontologies/ONTOTIOTOPE>
<https://reseau-cirm.hub.inrae.fr/>

p. 3

3

Extraction et normalisation de l'information scientifique

Intelligence artificielle et traitement automatique de la langue

Extraction d'entités complexes : termes longs et discontinus, grande variété de domaines et de nommage

Bacillus subtilis	Lives in	soil
Bacillus subtilis	Lives in	sourdough bread
Bacillus subtilis F91-69com	Lives in	clean room
Bacillus subtilis TW177	Lives in	clean room
Bacillus subtilis subsp. subtilis 80170	Lives in	cell
Bacillus subtilis strain 168	Lives in	soil
Bacillus subtilis strain 168	Lives in	experimental medium
Bacillus subtilis	Lives in	school

Extraction de relations : des relations longues distance et ambiguës entre des entités nombreuses

compound-producing endophytic bacterium HUB-1-47 isolated from **Basium rostratum**, **Juglans nigra** and **Ulmus**. The morphological, physicochemical characteristics of strain HUB-1-47 and its 16S rDNA sequences, and the antimicrobial compound were analyzed by gas chromatography-mass spectrometry were studied. We found that the morphology of HUB-1-47 was very similar to that of *Bacillus subtilis* but presented some differences in shape and cell size. Growth evaluation showed that the lowest, highest, and optimum growth temperatures of HUB-1-47 were 12, 47, and 31 degrees C, respectively, which were different from those of reference strains. The 16S rDNA sequence analysis revealed a high similarity of 99% to those of *B. subtilis*. Based on these analyses, we named strain HUB-1-47 as **B. subtilis subsp. virginiana** **S. P. Zhou, K. Zhao, J. Liu et W. X. Peng, subsp. nov.** This is the first report on the analysis of antimicrobial compounds from endophytic bacteria. Our study identified a new resource of antimicrobial compounds through endophytic bacteria fermentation.

Normalisation avec de grands référentiels : des milliers à des centaines de milliers de classes. Apprentissage automatique few- ou zero-shot

11428999, 15084811, ... *Listeria monocytogenes* Lives in immunodeficient person


Occurrence in text (taxon) *Listeria_monocytogenes*, *Lm*, *Listeria_monocytogene*, *L_monocytogenes*

Occurrence in text (habitat) patient with impaired immune function, immunocompromised host, immune - suppressed cancer patient

Full source text 11428999, 15084811, 10951886, 35689050, 7616977, 23019807, 8787708, 21976773, 30430390, 31837132, 1463824

Omnicrobe
21 nov 2024/ AG Hal / S. Schbath

Entraînement et évaluation : à l'aide du corpus annoté *Bacteria Biotope* dans des challenges internationaux (BioNLP-ST)



Extracting Information about microbial biotopes and phenotypes

p. 4

4

➤ Omnicrobe aujourd'hui

Workflow et application : Intégration d'un grand nombre de modules et de dépendances basée sur *SnakeMake* et application basée sur *PostgreSQL* et *Python/flask*

Mise à jour : tri-annuelle, gros volume de données hétérogènes.

Sources

Base doc.	Nb de doc.
PubMed	3 200 000
GenBank	500 000
DSMZ	133 298
CIRM	11 038

Informations extraites (non dédoublonnées)

Nb habitats	Nb microbes	Nb phenotypes	Nb uses	Relations habitats	Relations phenotypes	Relations uses
30 015 773	10 835 209	2 444 337	181 392	10 112 699	516 017	151 541

Nb relations microbe-habitat par source

- PubMed
- DSMZ
- CIRM-CFBP
- GenBank
- CIRM-BIA
- CIRM-Levures

Informations dédoublonnées

GenBank, 296.517
DSMZ, 53.616
PubMed, 628.173

Nb de relations de PubMed

45.408 14.169 628.173

- Taxo n- Habitat relations
- Taxo n- Phenotype relations
- Taxo n- Us er elations

INRAE
Omnicrobe
21 nov 2024/ AG Hal / S. Schbath

Dérozier S. et al., (2023). *PLoS ONE* 18(1): e0272473. P. 5

5

➤ Exemple d'utilisation

Nouveau "yaourt" à base de plantes fermentées

matière première → **choix du consortium microbien** → **expérimentation de fermentation** → **propriétés organoleptiques** → **produit final**

Omnicrobe : sélection de toutes les propriétés à la fois

Habitat : vegetable juice (soy, rice, oat beverage)
Usage : acidification activity
Phenotype : mesophilic or thermophilic
Source : collection INRAE CIRM-BIA

Identifie et relie des informations décentralisées et les unifie : améliore l'exploitabilité
gain de temps pour les chercheurs, notamment sur les sujets transversaux

Chaix E., et al., (2019) *Food Microbiology*, 10.1016/j.fm.2018.04.011

Omnicrobe
<https://omnicrobe.migale.inrae.fr/>

7 espèces à tester pour la production de yaourt

201 souches à tester
Test d'acidification : fermentation du jus de soja dans un milieu expérimental

Harté, O., et al. (2020) *Food Microbiology*.

6

➤ Perspectives

Nouvelles méthodes de TAL : intégrant de la connaissance du domaine (thèse d'A. Tang)

Exploitation des graphes de connaissances d'Omnicrobe pour inférer de nouvelles connaissances et améliorer la cohérence (thèse Z. Le Roux)

Nouveaux types d'information microbienne : molécules dégradées et produites critiques pour des enjeux en procédés de transformation : (thèse X. Zhu)

Extension du nombre et de la diversité des sources documentaires, notamment aux textes complets des articles

INRAE
Omnicrobe
21 nov 2024/ AG Hal / S. Schbath

p. 7

7

université PARIS-SACLAY **INRAE**

INSTITUT DATAIA Science des données, Intelligence & Sociétés

MoiAGE MATHÉMATIQUES ET INFORMATIQUES APPLIQUÉES AU GÉNOMIQUE À L'ENVIRONNEMENT

STLO Hélène Falentin
Stéphanie Deutsch

Labex DigiCosme université PARIS-SACLAY

Mouhamadou Ba
Robert Bossy
Louise Déleger
Sandra Dérozier
Valentin Loux
Claire Nédellec
Sophie Schbath

LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE Pierre Zweigenbaum

SPo UMR

Cirm

INRAE
Omnicrobe
21 nov 2024/ AG Hal / S. Schbath

p. 8

8

Recherche par l'espèce, ici *Bacillus subtilis*

Champs de saisie

Source text	Scientific name of taxon	Ontobiotope class	QPS	Source
100373	<i>Bacillus subtilis</i> subsp. <i>subtilis</i>	cell	yes	PubMed
10212408	<i>Bacillus subtilis</i>	Lives in school	yes	PubMed
10216858, 4321930, ...	<i>Bacillus subtilis</i> DSM 402	Lives in salt	yes	PubMed
10347866	<i>Bacillus subtilis</i> DSM 4515	Lives in cat	yes	PubMed
10347890, 10643273, ...	<i>Bacillus subtilis</i>	Lives in stomach	yes	PubMed
10369778, 18663522, ...	<i>Bacillus subtilis</i>	Lives in coelomic cavity	yes	PubMed

9

Recherche par l'habitat, ici *school*

Champs de saisie

Source text	Scientific name of taxon	Ontobiotope class	QPS	Source
100373	<i>Bacillus subtilis</i> subsp. <i>subtilis</i>	cell	yes	PubMed
10212408	<i>Bacillus subtilis</i>	Lives in school	yes	PubMed
10216858, 4321930, ...	<i>Bacillus subtilis</i> DSM 402	Lives in salt	yes	PubMed
10347866	<i>Bacillus subtilis</i> DSM 4515	Lives in cat	yes	PubMed
10347890, 10643273, ...	<i>Bacillus subtilis</i>	Lives in stomach	yes	PubMed
10369778, 18663522, ...	<i>Bacillus subtilis</i>	Lives in coelomic cavity	yes	PubMed

10

Retour au texte annoté pour visualiser la relation entre *B. subtilis* et *school*

Microorganisms

Microorganism	Abn.	Abc.
<i>Pseudomonas</i>	1	1
<i>Bacillus cereus</i>	2	1
<i>Micrococcus</i>	2	1
<i>Corynebacterium</i>	1	1
<i>Nocardopsis</i>	1	1
<i>Brevibacterium</i>	1	1
<i>Sphingomonas</i>	1	1
<i>Nocardia dassonvillei</i>	1	1
<i>Staphylococcus</i>	1	1
<i>Micrococcus</i>	2	1

Habitats

Habitat	Abn.	Abc.
dust	10	1
anaerobic environment	1	1
air	1	1
animal	6	1
school	5	1

Phenotypes

Phenotype	Abn.	Abc.
resistant	1	1

Abstract: A total of 316 bacterial strains, including psychrotrophs, mesophiles and thermophiles, were isolated and identified from indoor dusts in **schools**, children's day care centres and animal sheds. Several species which had not previously been reported from indoor environments were found: *Sphingomonas*, *Brevibacterium*, *Nocardopsis*, *Denococcus* and *Rhodococcus/Gordona*. A new psychrotrophic actinomycete genus was also found in animal sheds, representing a new undescribed polyphagous type and an unusual whole-cell fatty acid composition. The indoor dusts of animal sheds contained mainly the Gram-negative genera *Pseudomonas*, *Pantoea*, *Flavobacterium* and *Xanthomonas* early in the indoor feeding season, but changed to a composition dominated by *Bacillus*, *Micrococcus* and mesophilic and thermophilic actinomycetes towards the end of the season. The dust contained, and air-borne bacterial flora in **schools** and day care centres were dominated by Gram-positive bacilli and actinomycetes, notably *Bacillus cereus*, *Brevibacterium brevis*, *B. licheniformis*, *B. subtilis* and species of *Ambrobacter*, *Corynebacterium*, *Rhodococcus/Gordona*, *Nocardopsis* sp., *Denococcus*, *Staphylococcus* and *Micrococcus*. Indoor air and dust contained *Nocardia* spp., *Acetivibrio*, *Acetivibrio*, *Acetivibrio*, *Acetivibrio*, *Bacillus cereus* and *Nocardopsis dassonvillei* with the status of hazard group II. Indoor dusts of animal sheds contained eight different 3-hydroxy fatty acids, the 2-hydroxy fatty acid 14:0 and two 10-methyl fatty acids, whereas in dusts from **schools** and day care centres, these were below the detection level (< 3.5 mg·g⁻¹). The 3- and 2-hydroxy fatty acids could be assigned to one or more of the dust-contained culturable strains, but 10-methyl C14:0 was not present in any of the strains isolated. The dusts from **schools** and children's day care centres contained 0.2-0.3 ng of endotoxin·mg⁻¹ and 0.5-1.4 ng of beta-D-glucan·mg⁻¹, whereas the dusts from animal sheds contained more 0.3-41 ng·mg⁻¹ and 0.3-56 ng·mg⁻¹, respectively.

11