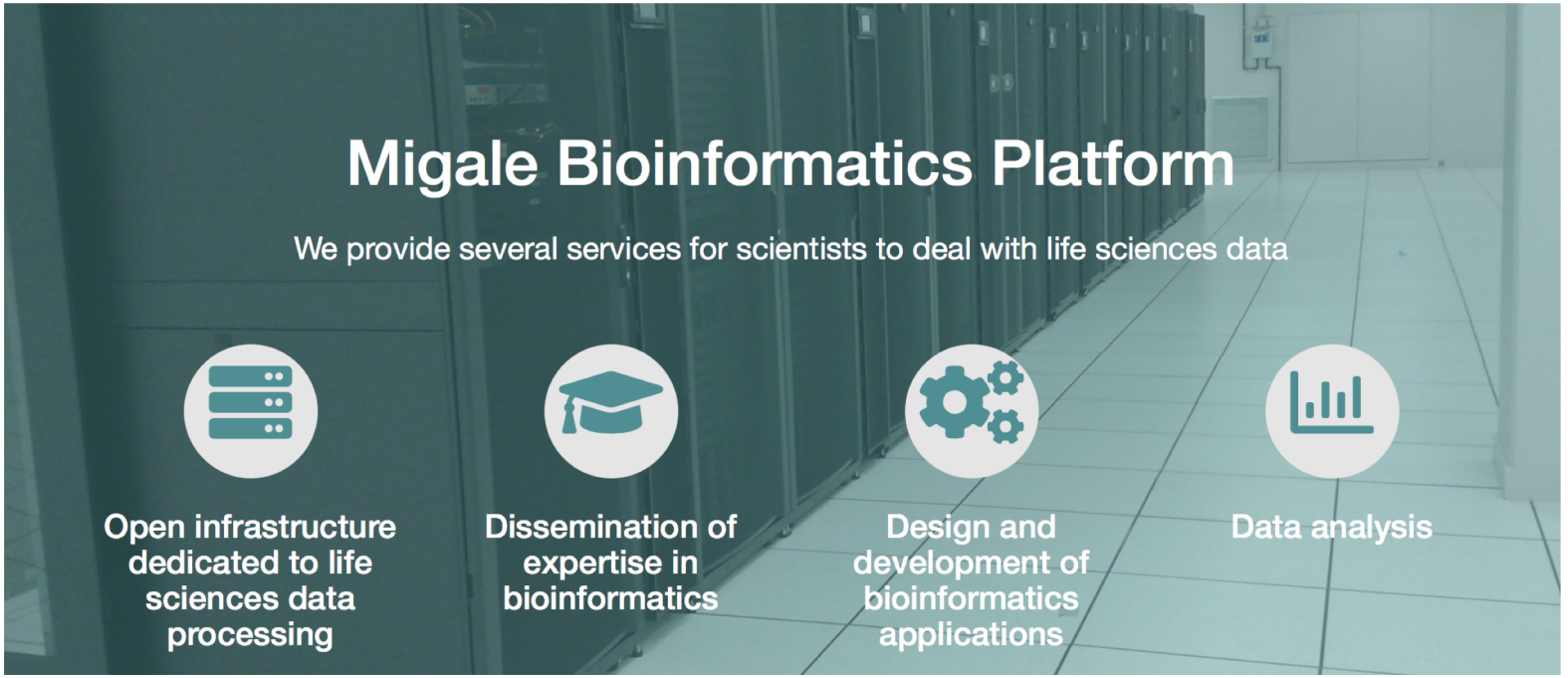
Quels microbes pour fabriquer un nouveau jus de lupin fermenté ? Le *text mining* à la rescousse !

Sophie Schbath,
MaIAGE, INRA, Univ. Paris Saclay, Jouy-en-Josas

Paris le 15/11/2019

La plateforme de bioinformatique migale

- Une **Infrastructure Scientifique Collective** de l'INRA, membre de
  - BioinfOmics : l'IR en bioinformatique de l'INRA
  - IFB : l'Institut Français de Bioinformatique
  - Elixir : le réseau européen pour l'information biologique



- et inscrite dans une démarche de Science Ouverte[1]

[1] http://institut.inra.fr/en/Overview/Documents/Charters/INRA-releases-official-open-access-guidelines

Un nouveau type de questions biologiques

Projet EnovFood (métaprogramme MEM, INRA)

Quelles bactéries pour fermenter un nouveau type d'aliments végétal, typiquement jus de lupin ou de soja ?

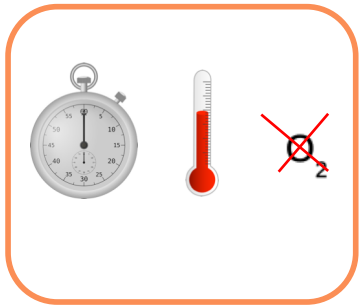
+ Disponibles dans la collection de souches du CIRM-BIA de Rennes.

Cirm
Bactéries d'intérêt alimentaire

Contexte : évolution des régimes alimentaires

Produire de nouveaux aliments, digestes, bons pour la santé, faciles à conserver

→ **La fermentation** : un procédé très intéressant …….



….. qui utilise un cocktail de microbes judicieusement choisis

Limite des approches classiques

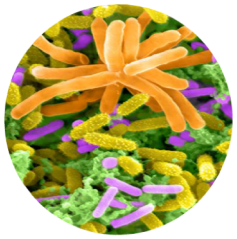
Approches classiques

- Tester chacune des 4000 souches de BIA , dans différentes conditions (milieux, températures, etc.) : trop coûteux et fastidieux
- Tester les souches de BIA en lien avec la fermentation : choix restreint
- Fouiller la littérature : difficile de chercher par mots-clés

Démarche visée

- **Explorer « automatiquement »** la littérature, mais aussi d'autres ressources textuelles (catalogues, bases de données génétiques ...)
- … cibler des connaissances sur les bactéries : habitats, leurs capacités à dégrader des molécules, etc.
- … pour trouver des bactéries capables de fermenter des jus végétaux ou présentes dans des aliments fermentés et « safe »
- Puis tester *in vivo* celles de BIA

Mon besoin en text mining plus général



Besoin plus large, issu d'équipes variées

- Quelles bactéries dans un habitat donné ?

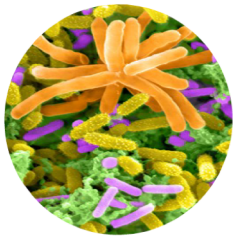
  Ex : Milieu salé, température, microbiote intestinal (sain/malade), etc.
- D'où provient cette bactérie trouvé dans cet aliment ?
- Quelles bactéries possèdent un phénotype donné ?

  Ex : dégradation de tel sucre, production de tel composé aromatique, etc.

Idéalement

- Avoir/construire une source d'information qui regroupe, pour une bactérie, les propriétés sur ses phénotypes, ses habitats, ses capacités de dégradation ou de production de molécules, etc.
- … que je puisse relier à d'autres informations (ex : génétiques),
- à partir des connaissances sous-exploitées contenues dans les publis et BdD

Objectif de la plateforme migale



Proposer un nouveau service basé sur du *text mining*

- permettant de lier les analyses bioinformatiques classiques (ex : étude de biodiversité par des approches de métagénomiques) à des informations extraites par *text mining*

Sans être spécialistes du *text mining*

- en nous appuyant sur l'équipe de recherche Bibliome pour le choix des outils spécialisés

→ **Florilège** : une application *pilote* co-construite de bout en bout sur un exemple

Utilisation de Florilège pour le jus de lupin

Home | Taxon lives in Habitat | Habitat may be inhabited by Taxon | Taxon exhibits Phenotype | Phenotype is exhibited by Taxon | Taxon studied for

- composite food
  - edible film
- food for particular diet
- liquid food
- plant product and primary derivative there
  - plant based drink
  - fruit and primary derivative thereof
  - garden vegetable and primary derivativ
  - grain and primary derivative thereof
  - legume seed and primary derivative the
    - bean and related product
    - lentil and related product
    - lupin and related product
      - lupin
      - lupin seed
    - pea and related product

Search relations by habitat

lupin and related product

TSV Download | Filter Selectio

6 relations for the habitat "lupin and related product"

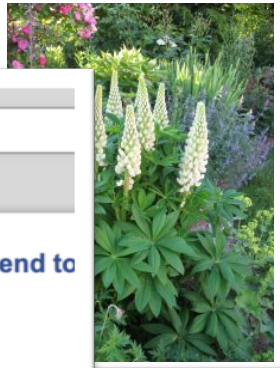
Source: PubMed, GenBank, CIRM, DSMZ

Taxon

QPS only

Apply

SOURCE TEXT	HABITAT	RELATION TYPE	TAXON	QPS	SOURCE
1790104	lupin	may be inhabited by	Lactobacillus acidophilus	✓	PubMed
1790104	lupin	may be inhabited by	Lactobacillus buchneri	✓	PubMed
1790104	lupin	may be inhabited	Lactobacillus	✓	PubMed



PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed | Advanced

Format: Abstract | Send to

Int J Food Microbiol. 1991 Dec;14(3-4):277-86.

Nutritional quality of lupine (Lupinus albus cv. Multolupa) as affected by lactic acid fermentation.

Camacho L[1], Sierra C, Marcus D, Guzmán E, Campos R, von Bäer D, Trugo L.

Author information

1 Institute of Nutrition and Food Technology (INTA), University of Chile, Santiago.

Abstract

The effects of selected NRRL strains of Lactobacillus acidophilus, L. buchneri, L. cellobiosus and L. fermentum upon oligosaccharide, phytate and alkaloid contents, as well as on the nutritive value of lupine, were investigated. Lupine was processed to a 12% total solids suspension, inoculated with 1% (v/v) cultures and fermented until a final desired pH of 4.5. L. acidophilus B-2092 and L. buchneri B-1837 growth was related to a significant sucrose breakdown and decreases of phytates, whereas L. acidophilus B-1910 and L. fermentum B-585 reduced the content of flatulence oligosaccharides. The activity of L. acidophilus B-1910 was particularly associated with lowering of alkaloids and increase of riboflavin. Lactic acid fermentation produced slight changes in lysine and

Home | Taxon I

- composite food
- edible film
- food for particular
- liquid food
- plant product and
  - plant based dri
  - fruit and primar
  - garden vegetab
  - grain and prima
  - legume seed a
    - bean and rel
    - lentil and rel
    - lupin and related product
      - lupin
      - lupin seed
    - pea and related product

xon studied for

SV load | Filter Selectio

QPS only | Apply

OURCE

ubMed

1790104	lupin	may be inhabited by	Lactobacillus buchneri	✓	PubMed
1790104	lupin	may be inhabited	Lactobacillus	✓	PubMed

Utilisation de Florilege pour le jus de lupin

- defrosted food
- ground food
- home-made food
- ⊞ mashed food
- ⊞ milled food
- prepared meat
- ⊟ preserved food
  - ⊞ canned food
  - concentrated food
  - ⊞ cooked food
  - cooled food
  - ⊞ dried food
  - ⊟ fermented food
    - fermented fish product
    - fermented meat
    - fermented dairy product
    - fermented seafood
    - fermented liquid
    - fermented cereal-based product
    - fermented plant-based food
  - ⊞ frozen food
  - ⊞ heat-preserved food

earch relations by habitat | fermented food | TSV Download | Filter Selection

14 relations for the habitat "fermented food

Source: PubMed, GenBank, CIRM, DSMZ | Taxon: Lactobacillus acidophilus | QPS only | Apply

CE TEXT	HABITAT	RELATION TYPE	TAXON	QPS	SOURCE
60422	kefir	may be inhabited by	Lactobacillus acidophilus	✓	PubMed
74688	semi soft cheese	may be inhabited by	Lactobacillus acidophilus	✓	PubMed
13050	sourdough	may be inhabited by	Lactobacillus acidophilus	✓	PubMed
68675	natto	may be inhabited by	Lactobacillus acidophilus	✓	PubMed
, 27112363	petit-suisse	may be inhabited by	Lactobacillus acidophilus	✓	PubMed
8275, 27157575	fermented food	may be inhabited by	Lactobacillus acidophilus	✓	PubMed
98684	fermented food	may be inhabited by	Lactobacillus acidophilus	✓	GenBank
	fermented milk	may be inhabited by	Lactobacillus acidophilus	✓	CIRM
14289, 18361739	dahi	may be inhabited by	Lactobacillus acidophilus	✓	PubMed
34161	kimchi	may be inhabited by	Lactobacillus acidophilus	✓	PubMed
, 28859276	kumis	may be inhabited by	Lactobacillus acidophilus	✓	PubMed
10289	salami	may be inhabited by	Lactobacillus acidophilus	✓	PubMed
94381	wine	may be inhabited by	Lactobacillus acidophilus	✓	PubMed
26237, 12822876	yogurt	may be inhabited by	Lactobacillus acidophilus	✓	PubMed

collection-cirmbia.fr/fiche.php?ncirm=445

INRA SCIENCE & IMPACT

Bactéries d'intérêt alimentaire

Centre de Cirm

Données de ba

Numéro CIRM BIA: 445

Souche type: non

N° collection déposante: CNRZ1295

Genre: Lactobacillus

Espèce: acidophilus — Lyophilisat: oui

Sous espèce: Not documented — Autres noms:

Whole Genome: 0

Autres collecti

CIP	DSMZ	ATCC	CNRZ	NCDO	NCFB
NR	NR	NR	NR	NR	NR

LMG	NCIMB	NCTC	CUETM	TL	INA
NR	NR	NR	NR	NR	NR

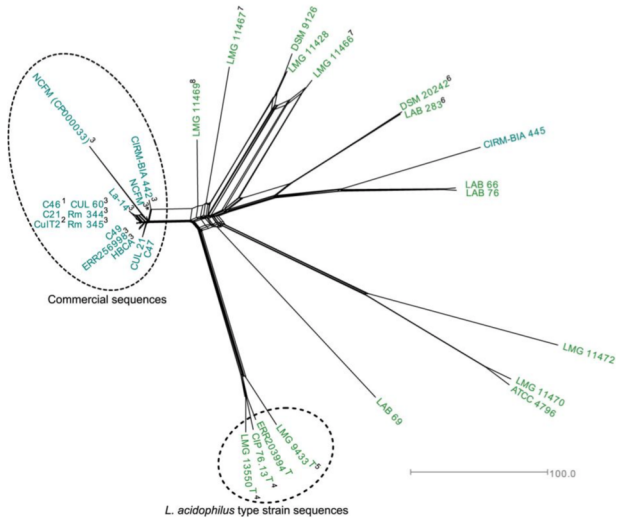
IL	AUTRE
NR	A6

Phénotypie

Atmosphère: Anaerobic - Milieu: MRS - T°C croissance: 43

Figure 2 : Whole genome MLST analysis of the infraspecies diversity of *L. acidophilus*.

From: The domestication of the probiotic bacterium *Lactobacillus acidophilus*



JGI GOLD — JGI HOME — LOG IN

Project Name	Lactobacillus acidophilus CIRM-BIA 445
Other Names	
Legacy ER Project ID	48101
Legacy GOLD ID	Gi0048101
NCBI BioProject Name	EMbaRC - Lactobacillus acidophilus CIRM-BIA 445
NCBI BioProject ID	200907
NCBI BioProject Accession	PRJEB1531
NCBI Locus Tag	
NCBI BioSample Accession	SAMEA2272655
PI	

prot.org

UniProt — UniProtKB — Lactobacillus acidophilus CIRM-BIA 445 — Advanced — Search

BLAST Align Retrieve/ID mapping Peptide search Help Contact

UniProtKB - V6BZS4 (V6BZS4_LACAI) — Basket

Display

Entry

Publications

Feature viewer

Feature table

None

Function

Names & Taxonomy

Subcell. location

Pathol./Biotech

PTM / Processing

BLAST Align Format Add to basket History

Help video Add a publication Feedback

Protein — Submitted name: **Proteolysis tag peptide encoded by tmRNA Lacto_acido_NCFM**

Gene — **tmRNA Lacto_acido_NCFM**

Organism — *Lactobacillus acidophilus CIRM-BIA 445*

Status — Unreviewed - Annotation score: ◉○○○○ - Protein predicted

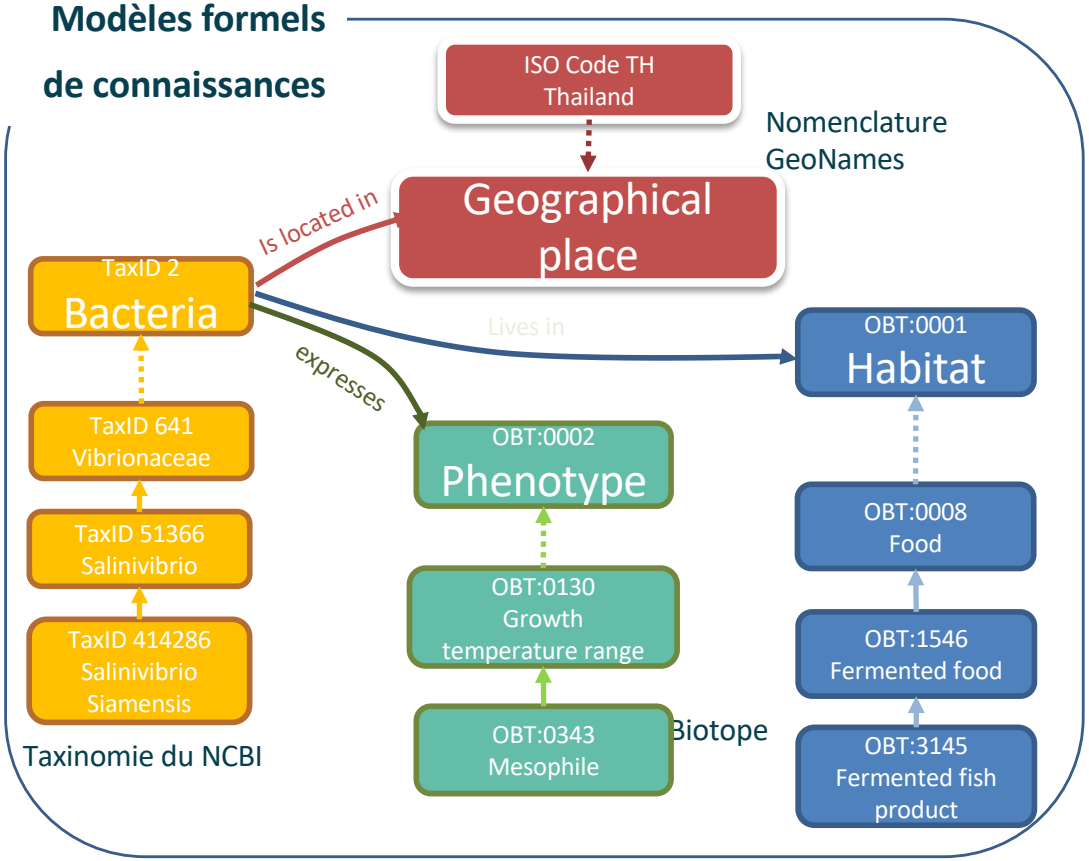
Names & Taxonomy

Protein names	*Submitted name:* Proteolysis tag peptide encoded by tmRNA Lacto_acido_NCFM — Imported
Gene names	*Name:***tmRNA Lacto_acido_NCFM** — Imported
Organism	Lactobacillus acidophilus CIRM-BIA 445 — Imported

INRA SCIENCE & IMPACT

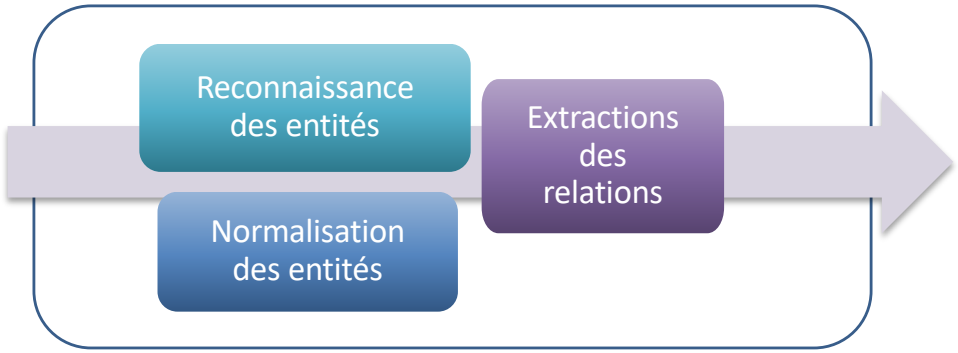
Un pipeline de TDM au cœur de Florilege

Modèles formels de connaissances

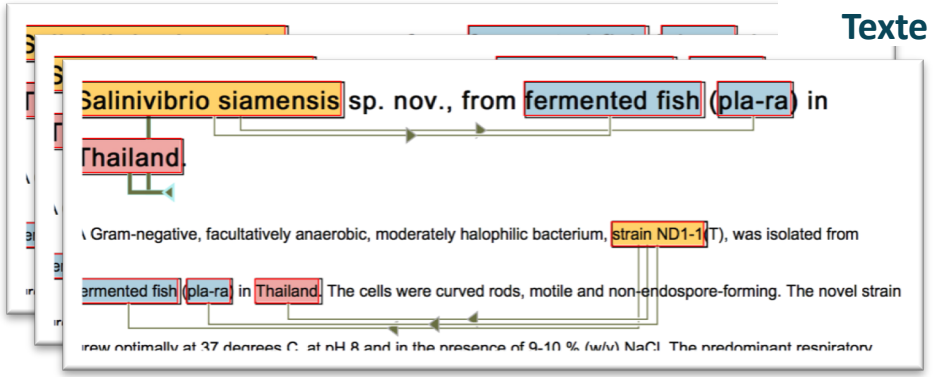


L'extraction d'information contribue aux principes

Extraction automatique d'information



Texte



Bilan du point de vue migale

Florilège

- une application cliente « à façon » que les biologistes se sont appropriés
  → [Falentin et al., Bageco'19]
- reposant sur une approche similaire à nos habitudes en bioinformatique : workflows d'outils exécutables, réutilisables, mutualisables, génériques

Migale : le bon endroit !

- Importance d'une application cliente **proche des besoins et adaptée aux biologistes.**
- **Plus-value :** lien entre les connaissances extraites des textes et les pipelines d'outils bioinformatiques classiques

  Ex : étude de la diversité microbienne (FROGS, etc.)

→ **Fort intérêt à inclure cette approche dans nos services**

Conclusions et perspectives

Objectif à court terme

Proposer un nouveau service aux utilisateurs (plus largement, BioinfOmics et IFB)

- Techniquement
  - Intégration du pipeline de text mining sur l'infrastructure de Migale
  - Aller vers plus d'automatisation et de généralisation

- Fort soutien de la part de l'INRA (recrutement + SDN)

À moyen terme : besoins immenses pour un passage à l'échelle

- Interne :
  - Tirer parti de la richesse de la représentation des connaissances extraites en investissant les techniques du **web sémantique**
- Externe :
  - Nécessité de nous reposer sur une **PF technique de TDM**
  - Besoin d'accompagnement de spécialistes du **TAL** (choix des outils)
  - Besoin d'outils et de partenaires **IST** (constitution des corpus)
  - Besoin de travailler avec les concepteurs **d'ontologies**

Conclusion : on y va mais pas tout seuls !

MaiAGE
Mathématiques et Informatique Appliquées du Génome à l'Environnement

STLO

openMINTED
Open Mining Infrastructure for Text & Data

INRA
SCIENCE & IMPACT

université
PARIS-SACLAY

http://migale.jouy.inra.fr/Florilege/

Catégoriser les termes du texte avec une ontologie - Un problème difficile

- Segmentation en phrases et en mots
- Filtrage de phrase et de documents
- Reconnaissance et normalisation des entités
- Étiquetage sémantique
- Extraction des relations

Exemple d'appariement

Terme du texte – concept de l'ontologie

Termes du texte

dairy
high hydrostatic pressure
hydrostatic pressure
hydrostatic
pressure
ival of the psychrotrophic organisms
psychrotrophic organisms
psychrotrophic
organisms
ultrahigh-temperature milk
ultrahigh-temperature
milk

"word embeddings" et alignement des vecteurs

appariement des têtes syntaxiques

appariement exact

Ontologie OntoBiotope

→ microbial habitat
→ food
→ animal product and primary derivative thereof
→ **milk and milk product**
→ butter
→ cheese
→ ice cream
→ **milk**
→ yogurt