Des microbes dans mon fromage ?

Valentin Loux & Sophie Schbath, MaIAGE, Inra, Jouy-en-Josas
Responsables de la plateforme bioinformatique Migale

Paris le 06/12/2018

La plateforme de bioinformatique migale

Missions :

- Mettre à disposition une **infrastructure de calcul scientifique** pour la génomique
- Diffuser un **savoir-faire en bioinformatique et biostatistique**
- Concevoir et développer des **applications**
- **Analyser** des données génomiques

FRANCE GÉNOMIQUE

ifb

CERTIFIED ISO 9001

Bioinformatique et Text/Data Mining (TDM)

Exemples de questions d'utilisateurs

- Puis-je avoir une liste des bactéries ayant été isolées dans un habitat donné ?
- Quelles sont les propriétés de ces bactéries ?
- D'où peut provenir cette bactérie que j'observe dans un milieu donné ?

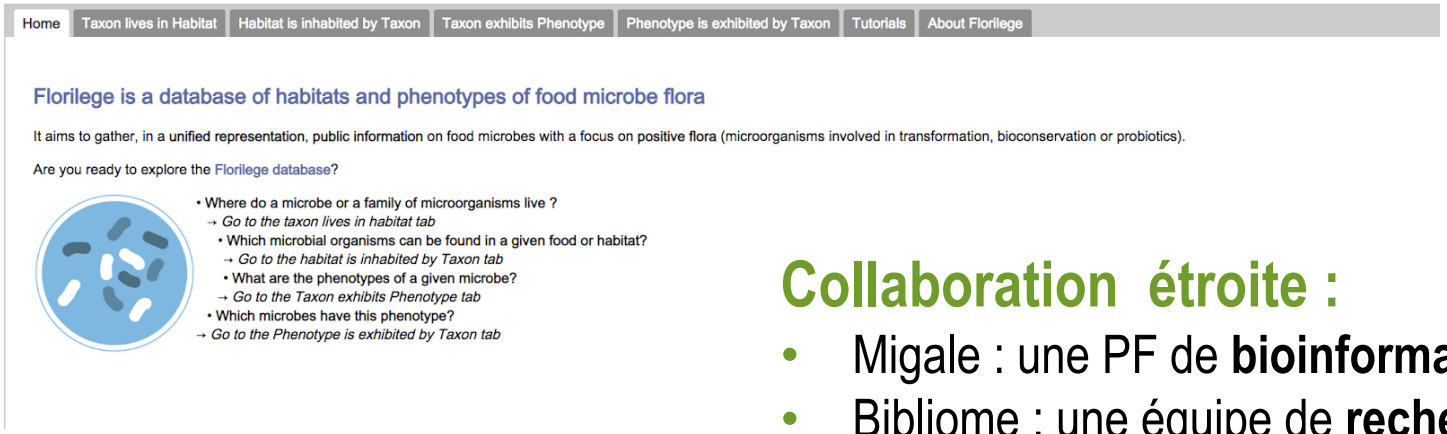
Motivation : Rôle positif des microorganismes dans les aliments

Protéolyse
Utilisation des sucres
Lipolyse
Arômes
Pigments
Vitamines
Biopréservation
Prévention de l'altération

Base de données Florilege (1/3)

Florilege est une base de données **d'habitats et de phénotypes microbiens extraits** de la littérature et de BD

~ 9000 vues en 2018



Collaboration étroite :

- Migale : une PF de **bioinformatique**
- Bibliome : une équipe de **recherche en TDM**
- Une dizaine d'équipes de **biologistes** du métaprogramme Inra MEM

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

Base de données Florilege (2/3)

Documents fouillés :

- Corpus Pubmed de 2,3 millions de résumés d'articles
- Banques ou ressources publiques :
  - BD génétiques (Genbank)
  - SI de centres de ressources biologiques (DSMZ, CIRM)

Intégration de données hétérogènes multi-sources

Outils de TDM développés :

- Reconnaissance d'entités nommées :
  - bactéries
  - habitats
  - phénotypes
- Appariement des termes avec une ontologie/taxonomie
- Extraction de relations via dépendances syntaxiques

} Verrous méthodologiques

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

Base de données Florilege (3/3)

Des documents à la base de données :

- Le traitement TDM est effectué par un workflow d'analyses déployé sur la **plateforme européenne OpenMinted**

- Résultats :
  - taxons : 8,4 millions
  - habitats : 18,5 millions
  - phénotypes : 1 million
  - 820 000 relations taxon ↔ habitat
  - 86 000 relations taxon ↔ phénotype

- L'ensemble des termes et des relations qui les lient est intégré dans une base de données relationnelle et accessible via une interface web : http://migale.jouy.inra.fr/Florilege

Exemple d'utilisation

Création d'un nouveau produit alimentaire

à partir de bactéries « QPS » (Qualified Presomption of Safety)

un fromage avec du sel

Recherche Florilege

onglet "Phenotype is exhibited by Taxon" → bactéries halotolérantes QPS Only

→ 9 Bactéries

Welcome | Taxon lives in Habitat | Habitat is inhabited by Taxon | Taxon exhibits Phenotype | **Phenotype is exhibited by Taxon**

Search relations by phenotype: halotolerant — TSV Download — Filter Selection

9 relations for the phenotype "halotolerant"

Source: OpenMinTeD, GenBank, CIRM, DSMZ — Taxon: — QPS only — Apply

SOURCE TEXT	PHENOTYPE	RELATION TYPE	TAXON	SOURCE
9327565	halotolerant	is exhibited by	Saccharomyces cerevisiae	OpenMinTeD
25039289	halotolerant	is exhibited by	Lactobacillus plantarum	OpenMinTeD
17897213, 25542205	halotolerant	is exhibited by	Bacillus pumilus	OpenMinTeD
15849794, 16467467, 17072537	halotolerant	is exhibited by	Debaryomyces hansenii	OpenMinTeD
18068256	halotolerant	is exhibited by	Lactococcus lactis	OpenMinTeD
12486459, 12557391, 21890005	halotolerant	is exhibited by	Bacillus subtilis	OpenMinTeD
18068256	halotolerant	is exhibited by	Leuconostoc lactis	OpenMinTeD
21664643	halotolerant	is exhibited by	Bacillus megaterium	OpenMinTeD
25561404, 16488097, 7646007	halotolerant	is exhibited by	Bacillus licheniformis	OpenMinTeD

1-9 of 9

Enjeux et Difficultés

Au delà des verrous méthodologiques de TDM :

- **Accès et traitement des documents** :
  - Récupérer les textes entiers (38% des publications ouvertes mais accès difficile)
  - Uniformisation des documents (format lisible par machine)

Ingénierie documentaire

- **Pluridisciplinarité de l'équipe-projet :**
  TDM/biologie/bioinformatique/ingénierie de la connaissance/documentaliste

- **Connecter ce type de service bioinformatique à une infrastructure de TDM :**
  OpenMinted à la française ?

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

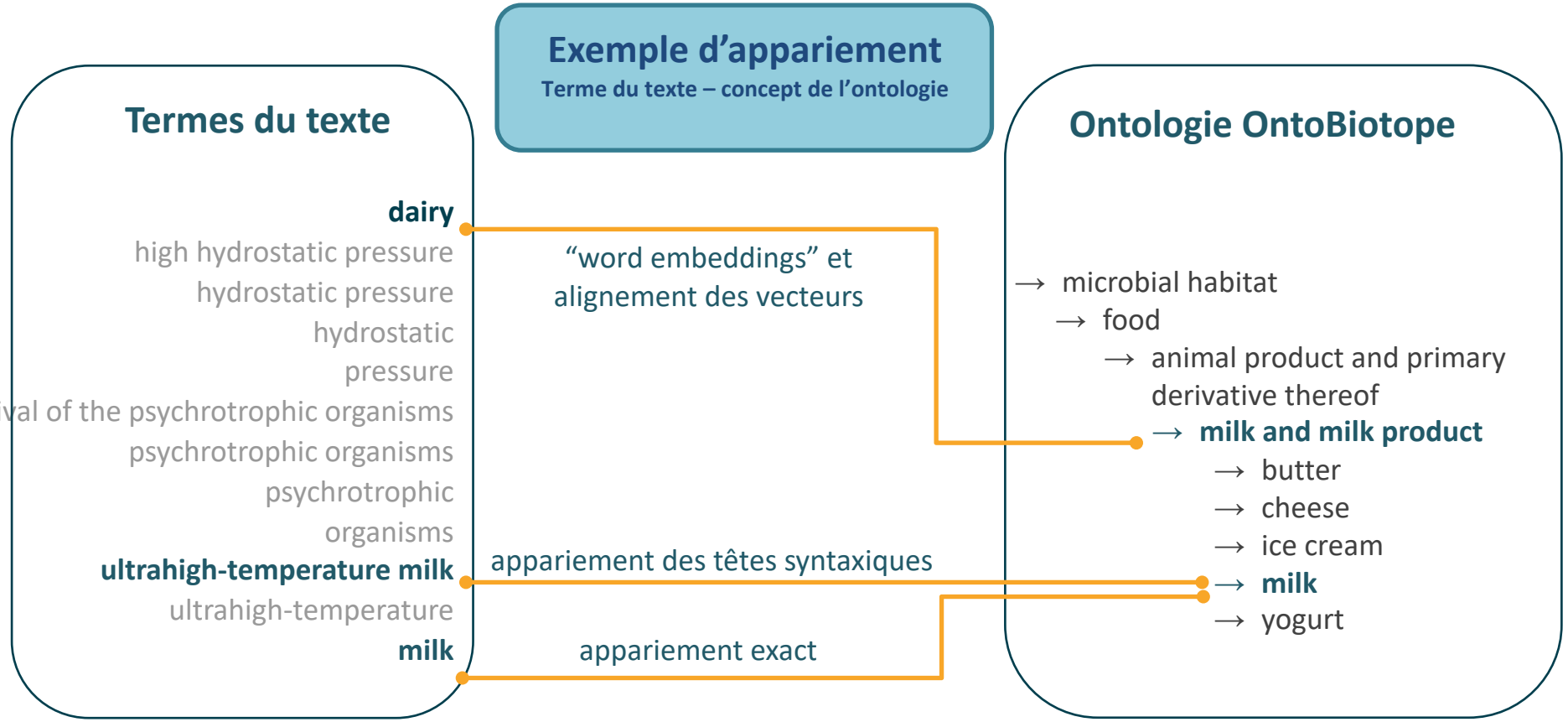
Segmentation en phrases et en mots

Filtrage de phrase et de documents

Reconnaissance et normalisation des entités

Étiquetage sémantique

Extraction des relations



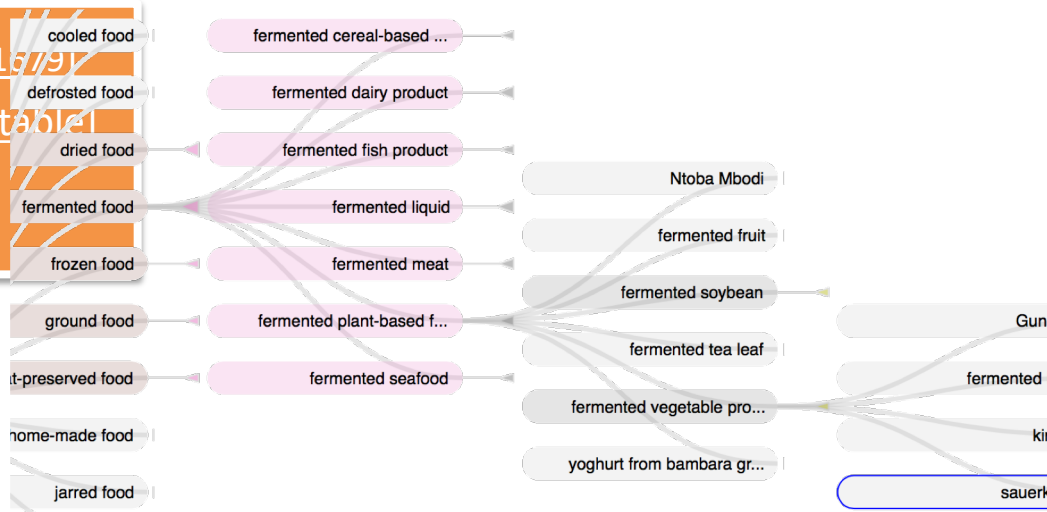
Textes

The addition of Propionibacteria to cabbage during sauerkraut production resulted in higher concentrations of vitamin B12 [Watanabe et al. (2014)]

Extraction d'information

Taxon *Propionibacteria* [taxid: 1678]
lives_in *sauerkraut* [fermented vegetable]
produces *vitamin B12* [cobolamine]

Catégories de l'ontologie OntoBiotope



Représentation structurée de l'information :

un formulaire dont les champs sont remplis par le texte et associés à des catégories