Text mining tools for extracting information about microbial biodiversity in food

Estelle Chaix, Louise Deléger, Robert Bossy and Claire Nédellec
firstname.lastname@inra.fr

Bibliome Team
MaIAGE, INRA, Université Paris-Saclay, 78590 Jouy-en-Josas, France

Text mining tools for extracting information about microbial biodiversity in food

Estelle Chaix, Louise Deléger, Robert Bossy and Claire Nédellec

Microbial ecosystems
Which microbes live in an environment?

Techniques:
- Culture-based methods
- Genetic
- Metagenomics
- Metatranscriptomics
- Shot gun sequencing
- and so on...

Helicobacter pylori
Mycobacterium avium
Escherichia coli ...

Legionella pneumophila
Yersinia pseudotuberculosis
Aeromonas hydrophila ...

Aspergillus flavus
Listeria seeligeri
Bacillus cereus ...

Properties of environment? Microbial Interaction?

Crossing between microbial species and habitats
Difficulty : Highly variable forms in text or genomics database (GOLD, SRA, GenBank)

“Bacteria of the genera Enterococcus and Lactobacillus and coliform bacteria were isolated from Dutch-type semi-hard cheese”

“Out of European red-s mare cheese samples of various types [...] 1.2% of the samples were contaminated with L. seeliger”

Habitat information is neither queryable nor comparable
Described at different levels of accuracy and not standardized

What is the cheese microflora?

“Geotrichum candidum strains isolated from a traditional Spanish goats’ milk cheese.”

“Escherichia coli O157:H7 isolated from raw beef, soft cheese and vegetables in Lima”

“Microbial ecology of Gorgonzola rinds and occurrence of different biotypes of Listeria monocytogenes.”

Text mining tools for extracting information about microbial biodiversity in food

Estelle Chaix, Louise Deléger, Robert Bossy and Claire Nédellec

Microbial spoilers in food 2017 - 28th - 30th June - Quimper

Text mining tools for extracting information about microbial biodiversity in food

Estelle Chaix, Louise Deléger, Robert Bossy and Claire Nédellec

Microbial spoilers in food 2017 - 28th - 30th June - Quimper
Text mining tools for extracting information about microbial biodiversity in food

Estelle Chaix, Louise Deléger, Robert Bossy and Claire Nédellec

Microbial spoilers in food 2017 - 28th - 30th June - Quimper

Classic search engine query

The query matches "cheese" and "microbe" but not "Camembert", "Roquefort" or "Listeria monocytogenes"

We propose a semantic search engine dedicated to microbial biodiversity in food.

Semantic search engine of microbial habitat in food

http://bibliome.jouy.inra.fr/demo/food/alvisir/webapi/search

Mini-link: https://frama.link/AlvisFood

Interpretation of the query

Cheese :
- American cheese
- Camembert
- Caerphilly cheese
- Cheddar cheese
- The Laughing Cow
- blue cheese
- camel's milk cheese
- compo cheese
- cinello cheese
- cottage cheese
- cream cheese

Has Aspergillus been isolated in cheese?

Results of the query: aspergillus cheese
Does Aspergillus lives in cheese?

Result of the query: *aspergillus* cheese

Does Aspergillus lives in cheese?

Result of the query: *aspergillus ~livesin cheese*

Our approach is to extract from text:
- "Microbe" and "Habitat" concepts
- Links between them

We use:
- AlvisNLP: Methods and tools for automatic extraction and analysis of biological text (i.e. Text Mining and Natural Language Processing)
- Machine learning methods trained with examples from microbiological and food domain experts
- Internal and external resources

AlvisFood Search Engine: > 100,000 references from PubMed
- Selected by MeSH terms
**Microbial entity detection**

NCBI taxonomy

- **Fungi**
  - Blastocladiomycota
  - Blastocladiales
  - Blastocladiaceae
  - Incertae sedis
  - environmental samples
  - uncultured Blastocladiomycota

- **Chytridiomycota**
  - Chytridiales
  - Cladophyllales
  - Gnomobiaceae
  - Tubulibacteriales
  - Meniscochlamydiales
  - Psychrotrichiales
  - Rhabdiophylloclaceae
  - Rhizopodiales
  - Spizellomycetaceae
  - unclassified Chytridiomycota
  - Chytridiomycetaceae incertae sedis
  - environmental samples
  - *Monosporobacteriales*
  - *Unclassified Monosporobacteriales*

**Food sub-categories of Ontobiotope ontology**

- From the EFSA classification
- Enrichment by microbial and food domains experts
- Formal indication that “Roquefort” is a “Cheese”
  - allows semantic search
- Our automatic AlvisNLP tools link groups of words from the text to an Ontobiotope category
  - achieve normalisation

**Habitat entity detection**

- Detection in text of nominal or adjectival groups
- Categorization of these groups with the **Ontobiotope ontology**
  - Formal and structured representation of microbial habitats
  - Partially reused in AlvisFoodSE

**Relationship between Microbe and Habitat**

- Extraction of ~livesin relationship
- Hard problems in automatic language processing and artificial intelligence
- Achieved by machine learning methods trained with annotated examples

What are the taxa living in food?

A query: `{taxon}* ~livesin food`

Results downloadable as table with occurrence counts. Displayed as facets.
To conclude

Our tools are pioneers in the field of text-mining for microbial biodiversity

Bibliome is a research team so:
- If you use AlvisFoodSE for your research, please cite us
- If you see an error, please send us an email, this will help us to improve our tools

On going work

- Ambiguous cases for automatic tools
- "Byssochlamys fulva and Neosartorya fischeri are heat-resistant fungi which are a concern to food industries"
- Automatic detection of microbial phenotypes
  
  i.e. halophile, thermophile, phototroph ...

Acknowledgments

INRA Ontobiotope and Florilège working groups
Food Microbiome project

Thank you for your attention