Facilitating the Screening of AMP Literature via Text Mining Strategies

Paula Jorge, Maria Olívia Pereira and Anália Lourenço*

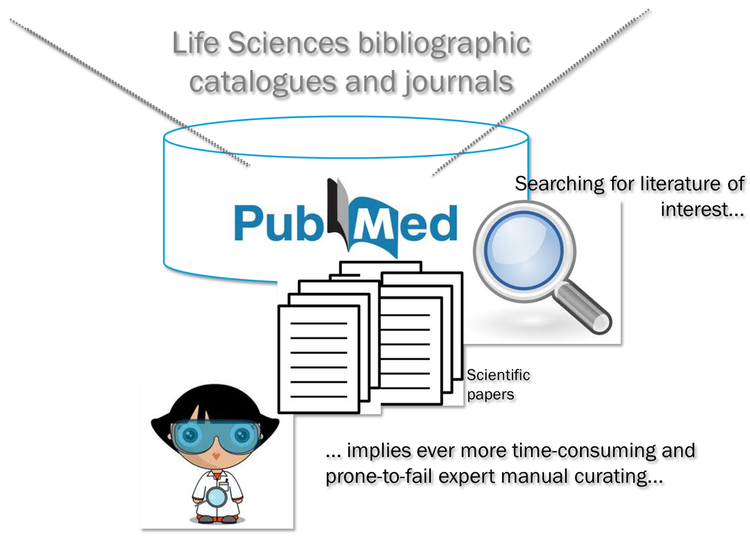
*IBB-Institute for Biotechnology and Bioengineering, Centre of Biological Engineering, Universidade do Minho,Campus de Gualtar 4710-057, Braga, Portugal

Introduction

Antimicrobial peptides (AMPs) are nowadays importantly recognized, mostly due to the raise of infections caused by multi-resistant pathogens, and have been hailed as a potential solution to the shortage of novel antibiotics, due to their general unspecific mechanism of action which impairs the acquiring of resistance.

Considerable efforts have been called to the study of antimicrobials with new mechanisms of actions, leading to a **proficiency of scientific publications and an emergence of specialized databases**. As the volume of related literature increases, it becomes urgent to help researchers and curators keeping up with it, especially by providing means for the **systematic screening and prioritization of literature of interest**.

Given the success reported in other biomedical domains, the integration of semi-automated mechanisms into AMP curation workflows would seem to be an ideal application for text mining.

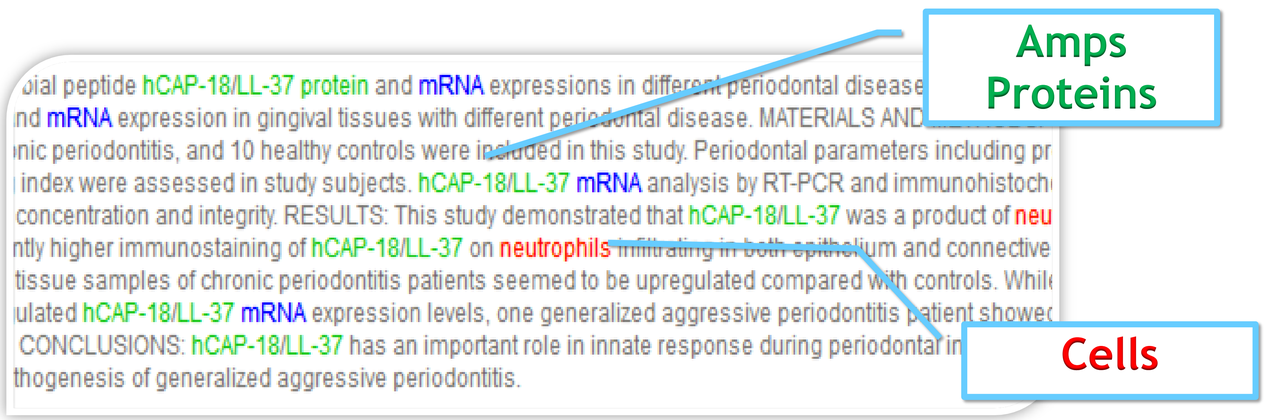


HOW CAN RESEARCHERS KEEP UP WITH NEW FINDINGS?
HOW CAN DATABASES BE KEPT UPDATED?

Text Mining:
screening + relevance assessment + information extraction

Aim

The goals of the work presented here were set into the development of a text mining prototype and the evaluation of its potential value in augmenting AMP curation throughput and efficiency.

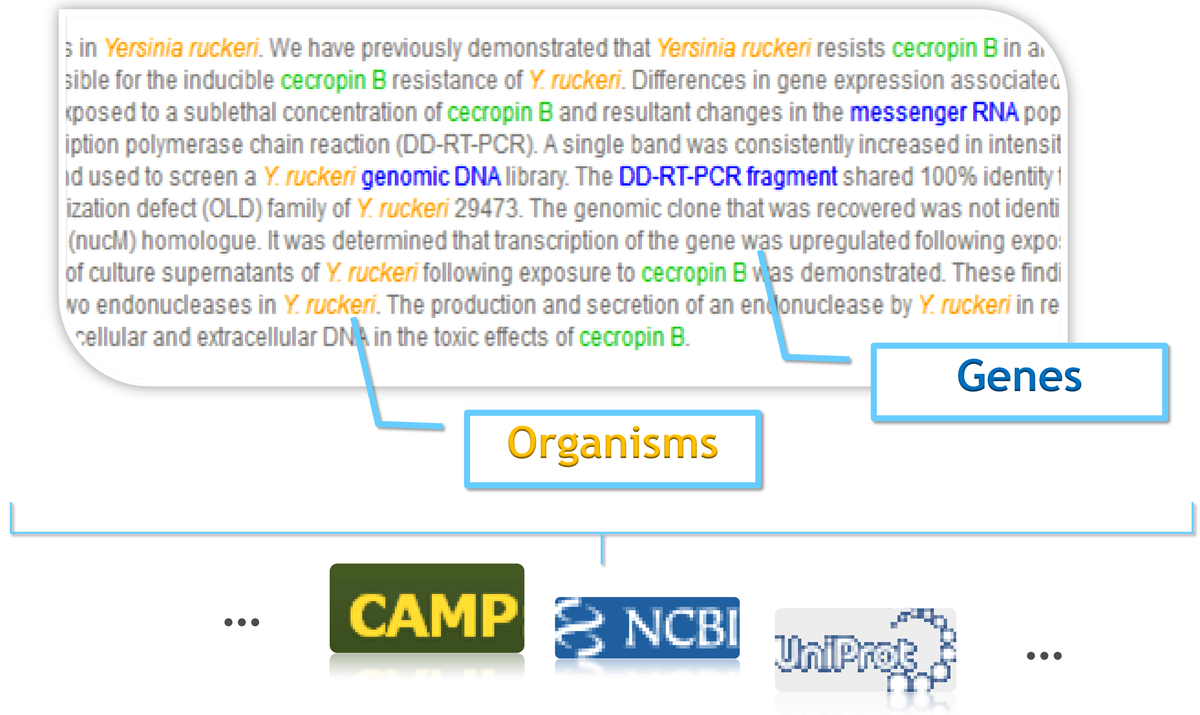


Methods

The text mining prototype was developed and evaluated using a corpus consisting of 1012 articles, published during 2011 and early 2012. The strategy to recognise biological entities of interest included:

- State-of-the-art **bio-recognisers** were used to tag **organisms** (LINNAEUS), **chemical entities** (OSCAR 4), and **genes**, **proteins** and **other biological entities** (ABNER).
- **AMP terms** were recognised using dictionary-based pattern matching. Vocabulary on AMPs derived from UniProt knowledge base, antibiotics lexicon from ARDB database and the **antibiotics** list in Wikipedia.
- An in-house lookup list of **textual clues on AMP bio-activities and mechanisms of action** was also included.

Article screening and ranking were based on statistical considerations. Especially, the number and diversity of unique concepts detected in the texts weighted by the relevance associate of the different biological categories and the degree of certainty associated to detection.



Examples of abstracts with highlighted and automatically annotated terms.

Preliminary Results and Ongoing Work

- **Preliminary prototype results are available at http://stardust.deb.uminho.pt/amps.**
- A list of articles ranked by year, pmid, title and journal is presented as proof of concept. Users will be able to order articles by these categories.
- To retrieve descriptions of AMP sources or targets, users are able to search using as a query the taxonomic name or common name of the organism of interest.
- Also, search for commons drugs and antibiotics present in the papers is available, as well as for genes and proteins.

Acknowledgments

The financial support from IBB-CEB, Fundação para a Ciência e Tecnologia (FCT) and European Community fund FEDER, through Program COMPETE, in the ambit of the FCT PTDC/SAU-SAP/113196/2009/FCOMP-01-0124-FEDER-016012 are gratefully acknowledged.

uminho BIOFILM GROUP

University of Minho
School of Engineering

IBB