**Automated bioactive peptide literature search and MBPDB updates**

**Summary:** Develop a system that searches the biomedical literature for milk derived bioactive peptides to be incorporated into the MBPDB. The system will allow researchers to periodically query Uniport and PubMed for bioactive peptides then extract their biological function, sequence information and reference’s. This list of peptides, functions and references will be compared with the existing MBPDB database to identify novel peptides. The bioactivity, legitimacy, and compatibility of the peptides absent from the MBPDB database will then be assessed by the researchers with the assistance of AI derived text extraction and summarization tools. Peptides that meet the requirements for inclusion in the MBPDB can then be added to the database allowing for additional macro analysis and a more comprehensive database.

**Primary Objectives:**

* Develop a tool to periodically search the literature (Uniport and PubMed) for bioactive peptides absent from the MBPDB
  + Uniport: query based on protein ID of list of known milk proteins
  + PubMed: Period searches of new publications based on previously established key words
* Automate the extraction and transformation of the peptide information to save researcher’s time
* Implement a system/process that is standardized and allows for regular database updates
* Incorporate AI text extraction and compression tools for additional time saving
* Contribute a significant number of novel peptides from the initial search to allow for additional macro analysis of the database, providing insight for a new publication
* Ensure the search and text extraction/comprehension process is robust and documented to allow for incorporation into a manuscript
* Publish an article covering the new search process and analysis of the newly added peptides
* Lay the foundation for additional grants funding further work

**Secondary Objectives:**

* Incorporate the tool into the user interface of the MBPDB web application for ease of use by future researchers
* Develop a customized large language model (LLM) that is trained on the bioactive milk peptide literature to improve the accuracy of text comprehension
* Allow for natural language queries to this customized LLM improve the usability and insight from the MBPDB
* Search PubMed using the LLMs text comprehension ability to identify articles of interest as an autonomous AI agent instead of simple key word searches
* Search PubMed using each peptide in the MBPDB as a key word to find new references and bioactivity for existing peptides
* Incorporation of post transcriptional modifications peptides

**Methodology:**

**Literature Search**

Develop script that connects to PubMed and Uniport through an API for automated searches

üSearch Uniport given a list milk Proteins and download meta data for each protein

üStart with Dave’s list of cow and human milk proteins

û Expand the protein list to cover additional species and update the cow and human milk protein list

û Search PubMed given a list of key words and identify articles of interest

üA list of key words exists in the first MBPDB publication

û Use each peptide in the database as a key word to identify new references/ biological functions

**Text Extraction**

Extraction the relevant meta data associated with each peptide and references from the literature search

üExtract peptide, bioactivity, and reference metadata through API calls from Uniport.

üAssociated peptide, bioactivity, and reference data for each Uniport identified peptides

üExtract reference metadata through API calls from PubMed

û Extract peptide and bioactivity information from articles or abstracts identified in PubMed search

û Extract ic50, ptm or microbial inhibition information from the text for either Uniport or PubMed data

This table shows the fields required to add a peptide to MBPDB and what is currently being extracted from Uniport search script

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Protein ID | peptide | function | Additional details | ic50 | Inhibition type | Inhibited micro-organisms | ptm | title | authors | abstract | doi |
| ü | ü | û | û | û | û | û | û | ü | ü | ü | ü |

**Text Analysis**

Have a system to assess the legitimacy of a peptides biological function claim, how to better classify it into pre-existing MBPDB bioactivity functions then re-classify these peptides biological function.

û Establish the criteria for the legitimacy of a bioactivity peptide functionality claim

û Establish the criteria for what constitutes bioactivity of a peptide compared to a non-bioactive function

û Incorporate a text comprehension AI system to classify the bioactivity of a peptide into existing functions found in the MBPDB

û For peptides with bioactivity functions absent from the MBPDB create simple broad functional classifications for these peptides

û Figure out what information to populate the additional details column with (likely the original bioactivity highly technical claim found in Uniport)

û Analyze the abstract or peptide’s manuscript to extract the ic50, microbial inhibition and type for the text

û Provide sufficient documentation for this analysis to be audited by a human researcher

**Compile Data**

After analysis of the text compile the results for review by a researcher and upload to the MBPDB

û Develop a script the joins and formats the different data sources into two tables

* MBPDB upload .TSV file formatted to the table above
* A table that provides the documentation for any decisions made for reclassification of a bioactivity function or other text analysis

**Analyze Data**

Preform analysis on the newly added peptides to highlight discoveries from the additions

û Preform macro analysis on the updated collection of peptides in the MBPDB

û Address questions that can now be answered with the inclusion of the new peptides

û For example can we better analyze the antimicrobial properties of these peptides

**Manuscript Preparation**

Write and submit manuscript on the database updates

û Discuss and document

* The system for extracting peptides from the literature
* Incorporation of any AI tools
* Expanding the scope of the MBPDB to be more comprehensive
* The new peptides added, their protein origins, functionality etc
* Macro analysis of the entire peptide database
* Website updates for functionality and ease of use etc

**Update the MBPDB**

Upload the new peptides to the database and deploy the new image to the cloud

**Current Problems:**

* Find a comprehensive list of milk proteins from human, cow, sheep, goat, pig, yak, rabbit, donkey, camel, buffalo which are the species in the MBPDB
* Establish the criteria for article inclusion in the manuscript. Example questions to consider the eligibility of an article:
  + Does the article have to directly reference milk?
  + What constitutes a valid bioactive function?
  + What are the reference requirements for a biomedical claim of a peptide?
* How to extract the peptide and functionality information from the abstracts/papers found in the PubMed Search
* How to bulk download papers from the PubMed Search in an automated legitimate manor