

AI tools for large-scale text mining PubMed

Zichen Yang

University of Liverpool

career@ihack.uk

Overview

Background

- **Source:** 7M+ open-access publications from  PubMed Central® in unstructured text
- **Goal:** Databases like  VEuPathDB require structured, high-quality data (e.g.,  Gene Ontology annotations) to be useful for researchers
- **Bottleneck:** Manual curation from the literature is labor-intensive & slow

Question

How can we build an AI pipeline to propose accurate specie–gene–GO annotations for PubMed articles to accelerate the curation process?

MCP Servers

- **MCP (Model Context Protocol)** is an **open** standard for LLM to interact with external tools and data sources, such as web searching, uploaded files, etc.
 - It is supported in some AI chat apps (e.g. Claude Desktop, Cherry Studio)
- MCP follows client-server architecture, where the client is the LLM and the server provides access to external resources.
- We developed three MCP servers ( PeronGH/biomcp):
 - **PubMed** provides access to open-access articles in  PubMed Central®
 - **PlasmoDB** provides the ability to search PlasmoDB (part of  VEuPathDB)
 - **QuickGO** allows querying of Gene Ontology annotations

Agent (MCP Client)

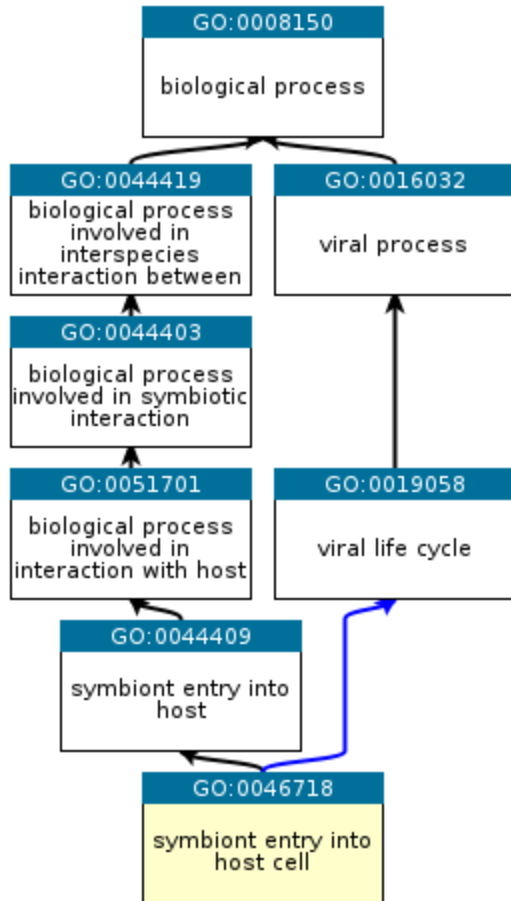
- An **AI agent** is a system that can **autonomously** perform tasks by interacting with the environment.
- We developed a single-agent system using Google Agent Development Kit
- Workflow:
 - i. Takes a PMID (ID of a PubMed article) as the input
 - ii. Queries the MCP servers for relevant information
 - Fetches full text of the article using PubMed
 - Identifies species and genes mentioned in the article using PlasmoDB
 - Searches QuickGO for relevant Gene Ontology annotations
 - iii. Outputs the gathered information in a structured format

Results

Compared to human curators, our AI agent shows:

- High recall on gene identification
 - Is usually able to identify all relevant genes mentioned in the article
 - Sometimes identifies additional genes (not validated)
- Reasonable but different  GO annotations
 - Sometimes proposes more specific GO terms than human curators
 - Sometimes proposes related GO terms but in a different aspect
 - Unable to provide metrics (hard to decide which prediction is true)

Case Study



QuickGO - <https://www.ebi.ac.uk/QuickGO>

Left

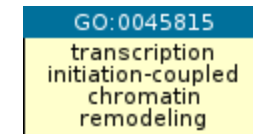
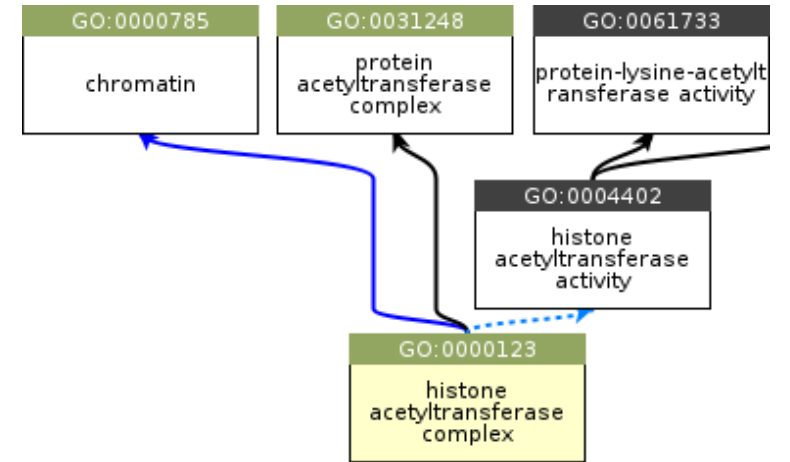
PMID: 30794532, Gene: *PDEβ*

- GO:0044409 (Human)
- GO:0046718 (AI)

Right

PMID: 34403450, Gene: *GCN5*

- GO:0000123 (Human)
- GO:0045815 (AI)



Process

Function

Component

Discussion

- Our MCP servers show potential for usage beyond this specific AI agent
- Our AI agent can identify species and genes but struggles with proposing accurate GO annotations
 - Future works:
 - Refine the agent's prompt with feedback from biologists
 - Adopt more advanced agent architectures, such as multi-agent systems.

Credits

- Antony McCabe for supervising the project
- Daniel Warren and Jordan Tzvetkov for providing feedback
- HPC team at the University of Liverpool for providing access to Barkla2 HPC