Functional Annotation of PATRIC using Gene Ontology

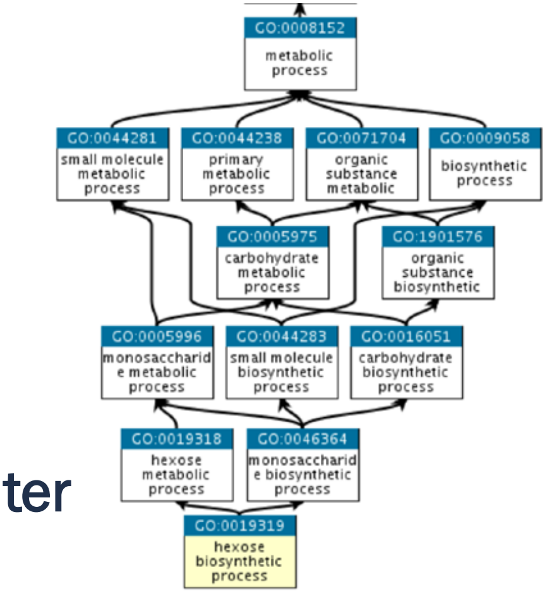
Jainam Modh

Dr. Andrew Warren, Dustin Machi, and Joseph Outten

Background

Gene Ontology (GO)

- A network of biological classes that describe the current best interpretation of the "universe" of biology
- A directed acyclic graph representation of protein function separated into three aspects:
  - Molecular Function
  - Cellular Component
  - Biological Processes


GO:0008152
metabolic process
GO:0044281
small molecule metabolic process
GO:0044238
primary metabolic process
GO:0071704
organic substance metabolic
GO:0009058
biosynthetic process
GO:0005975
carbohydrate metabolic process
GO:1901576
organic substance biosynthetic
GO:0005996
monosaccharide metabolic process
GO:0044283
small molecule biosynthetic process
GO:0016051
carbohydrate biosynthetic process
GO:0019318
hexose metabolic process
GO:0046364
monosaccharide biosynthetic process
GO:0019319
hexose biosynthetic process

PATRIC

- PAthosystems Resource Integration Center
- Database for functional annotation of bacterial and viral genomes used commonly in infectious disease and antibiotic resistance research.
- Proteins/coding sequences are poorly annotated and not updated with GO terms.

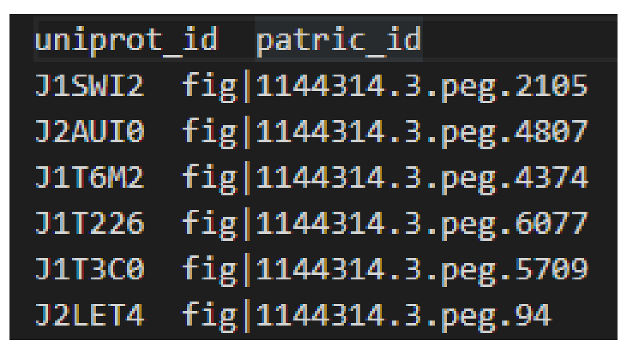
Current Work

Extraction of md5 hashes and IDs from PATRIC

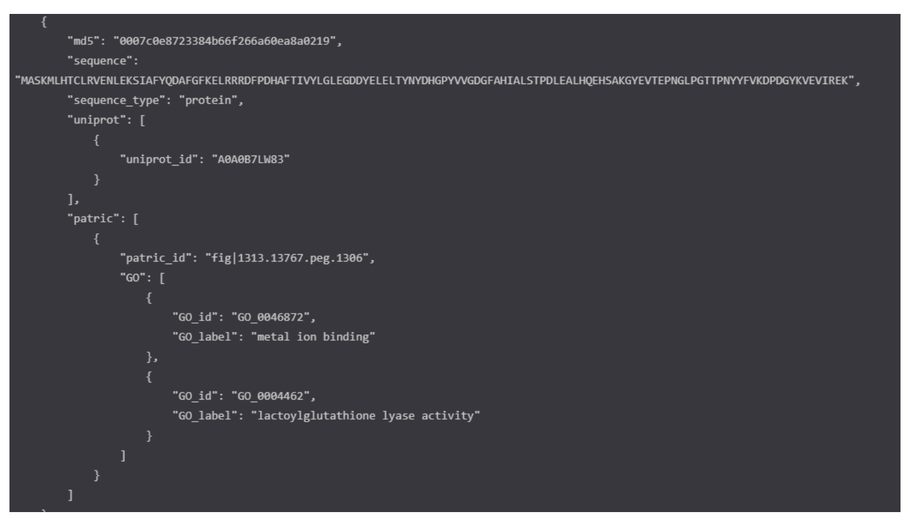
- Use the PATRIC Command Line Interface (CLI) to extract all coding sequences from 459955 bacterial and viral genomes in smaller batches of 250 genomes each.
- Each amino-acid sequence has its own md5 hash that is associated with multiple PATRIC IDs from different bacterial genomes.
- Extract other relational information including genus-specific protein families (PLfams) and cross-genus protein families (PGfams).

Query UniprotKB for GO Terms

- Use md5 hash and the taxonomic identifier from PATRIC to query the UniprotKB database through the SPARQL endpoint.
- Construct a table of link-outs which connect the Uniprot IDs to PATRIC IDs in a many to many relationship.
- Construct a JSONhFasta file that assigns GO Terms to each protein sequence and PATRIC ID in the PATRIC database.



uniprot_id	patric_id
J1SWI2	fig|1144314.3.peg.2105
J2AUI0	fig|1144314.3.peg.4807
J1T6M2	fig|1144314.3.peg.4374
J1T226	fig|1144314.3.peg.6077
J1T3C0	fig|1144314.3.peg.5709
J2LET4	fig|1144314.3.peg.94



```
    {
        "md5": "0007c0e8723384b66f266a60ea8a0219",
        "sequence":
"MASKMLHTCLRVENLEKSIAFYQDAFGFKELRRRDFPDHAFTIVYLGLEGDDYELELTYNYDHGPYVVGDGFAHIALSTPDLEALHQEHSAKGYEVTEPNGLPGTTPNYYFVKDPDGYKVEVIREK",
        "sequence_type": "protein",
        "uniprot": [
            {
                "uniprot_id": "A0A0B7LW83"
            }
        ],
        "patric": [
            {
                "patric_id": "fig|1313.13767.peg.1306",
                "GO": [
                    {
                        "GO_id": "GO_0046872",
                        "GO_label": "metal ion binding"
                    },
                    {
                        "GO_id": "GO_0004462",
                        "GO_label": "lactoylglutathione lyase activity"
                    }
                ]
            }
        ]
```

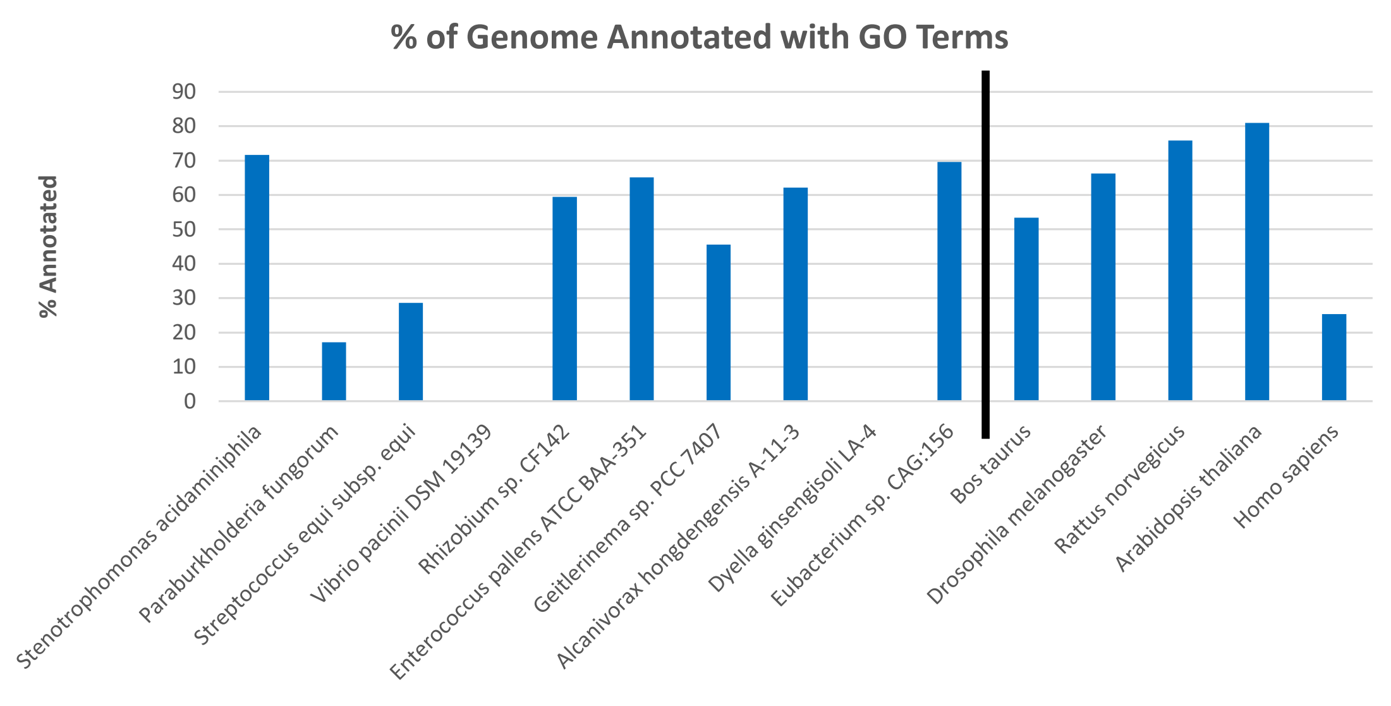


Figure #2. Analysis of PATRIC GO Annotations

Project Goals

Overall Goal: Automated Function Prediction using Gene Ontology

- Predict the GO terms for an unknown coding sequence using the existing network of GO graphs.

Current Goals: Improve GO Annotation of Bacterial Genomes

- Update crosslinks between the proteins IDs in the UniprotKB and PATRIC databases.
  - Uniprot KnowledgeBase is the largest repository of protein sequences and annotations. It attempts to gather information from multiple biological databases for centralized access and storage.
  - Current connections between the two databases are outdated and incomplete.
- Annotate PATRIC with GO Terms queried from UniprotKB
- Analyze bacterial annotations for expandable GO Terms that could have more children

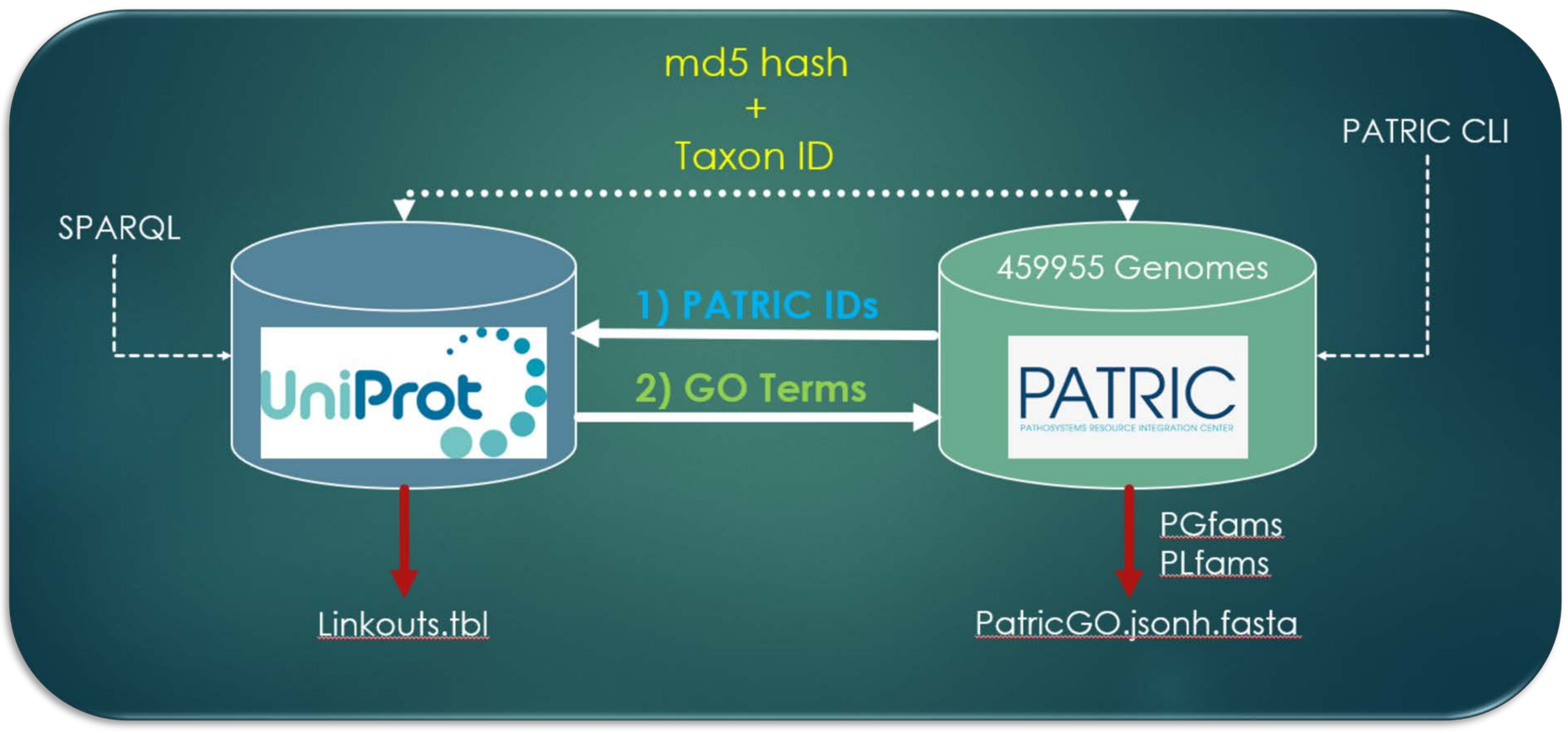


Figure #1. Overall database query structure for updating PATRIC and UniProt.

Future Work

Resolve Issues with Larger Query Sizes

- Both databases are extremely large and expensive to query leading to memory errors and faulty server endpoints.
  - Extracting a single batch of 250 genomes using PATRIC CLI takes 40 min.
  - Querying those 250 genomes for GO Terms through Uniprot's SPARQL endpoint takes more than 20 hours.
- Proposed solution is to locally download the entire Uniprot database to avoid queries to remote servers.

Construct Automated Function Prediction Model for Bacterial Proteins

- Index through the JSONhFasta file to obtain sequences and GO graphs.
- Use DIAMOND, an all-to-all BLAST tool, to find clusters of similar sequences and create a weighted network of GO graphs.
- Determine the distribution of each possible GO Term in the network.
- Predict the probability for a new protein sequence to have a set of GO Terms based on the existing distribution of GO Terms and network of protein sequences.

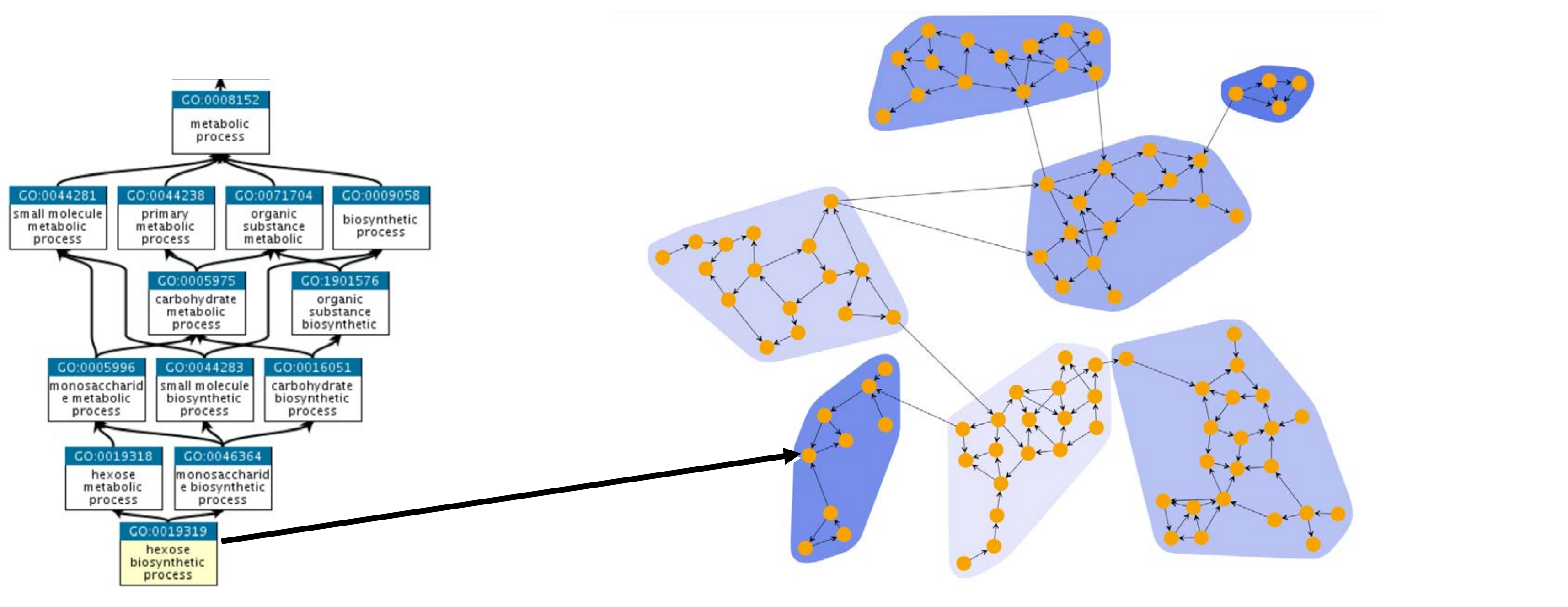

GO:0008152
metabolic process
GO:0044281
small molecule metabolic process
GO:0044238
primary metabolic process
GO:0071704
organic substance metabolic
GO:0009058
biosynthetic process
GO:0005975
carbohydrate metabolic process
GO:1901576
organic substance biosynthetic
GO:0005996
monosaccharide metabolic process
GO:0044283
small molecule biosynthetic process
GO:0016051
carbohydrate biosynthetic process
GO:0019318
hexose metabolic process
GO:0046364
monosaccharide biosynthetic process
GO:0019319
hexose biosynthetic process

Figure #3. Automated function prediction model using a combination of DIAMOND and Gene Ontology

References

- James J Davis et al. The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities, Nucleic Acids Research, Volume 48, Issue D1, 08 January 2020, Pages D606–D612, https://doi.org/10.1093/nar/gkz943
- BuchfinkB, Reuter K, Drost HG, "Sensitive protein alignments at tree-of-life scale using DIAMOND", *Nature Methods* 18, 366–368 (2021). doi:10.1038/s41592-021-01101-x
- Shuwei Yao, Ronghui You, Shaojun Wang, Yi Xiong, Xiaodi Huang, Shanfeng Zhu, NetGO 2.0: improving large-scale protein function prediction with massive sequence, text, domain, family and network information, Nucleic Acids Research, Volume 49, Issue W1, 2 July 2021, Pages W469–W475, https://doi.org/10.1093/nar/gkab398

Computing
for Global Challenges

University of Virginia
BIOCOMPLEXITY INSTITUTE