SocialGene: Large Scale Analyses of Protein Similarity for Microbial Drug Discovery

Chase M. Clark, Jason C. Kwan; Division of Pharmaceutical Sciences; School of Pharmacy; University of Wisconsin–Madison

twitter.com/ChasingMicrobes

linkedin.com/in/hirechase

chasemc.github.io
Resume and CV

Summary

The annotation of function to genes is commonly achieved by the relation of sequence similarity to previously characterized nucleotide or protein sequences. In genomic-based microbial drug discovery this concept extends into the theory that homologous proteins and biosynthetic gene clusters (BGCs) will produce similar natural products (NP).To help find proteins of similar function, especially where protein sequence divergence is high, I have created a unified collection of tools (python package, nextflow pipeline, graph database, user interface), called SocialGene. SocialGene attempts to overcome some of the inherent difficulties of searching through large numbers of genomes, while retaining the flexibility of having no inbuilt notion of what a BGC should look like. Precomputing pHMM annotations of non-redundant proteins facilitates searching by domain homology and the discovery of distantly related proteins which may be missed by tools that rely on sequence alignment alone, e.g. BLASTp. Our planned use cases are broad and include discovering BGCs encoding new chemical analogs; strains to screen for increased BGC expression; suitable hosts for heterologous expression; finding BGCs across fragmented genomes; and performing targeted and co-occurring enzyme domains (CO-ED) analyses.

Protein "similarity" isn't a simple concept...

Our lab has a special interest in endosymbiotic bacteria which often have genomes and genes that have undergone significant evolutionary changes compared to free-living relatives. Finding homologous proteins to those found in these endosymbiotic bacteria is still a challenge for us, especially against large cutom sets of genomes and input BGCs.

While there are numerous ways to determine protein functional similarity, each with their own strengths and weaknesses, two of the most common are sequence alignment and Hidden Markov Model (HMM) annotation.

Sequence Alignment — diamond

- Faster
- Substitution matrices model likelihood of AA substitution
- Less sensitive, limited for remote homology ("twighlight zone")

HMM annotation — HMMER — HH-suite

- Slower
- pHMMs model AA substitution, insertions and deletions based on multiple sequence alignments
- More sensitive (better at finding distant homologs)

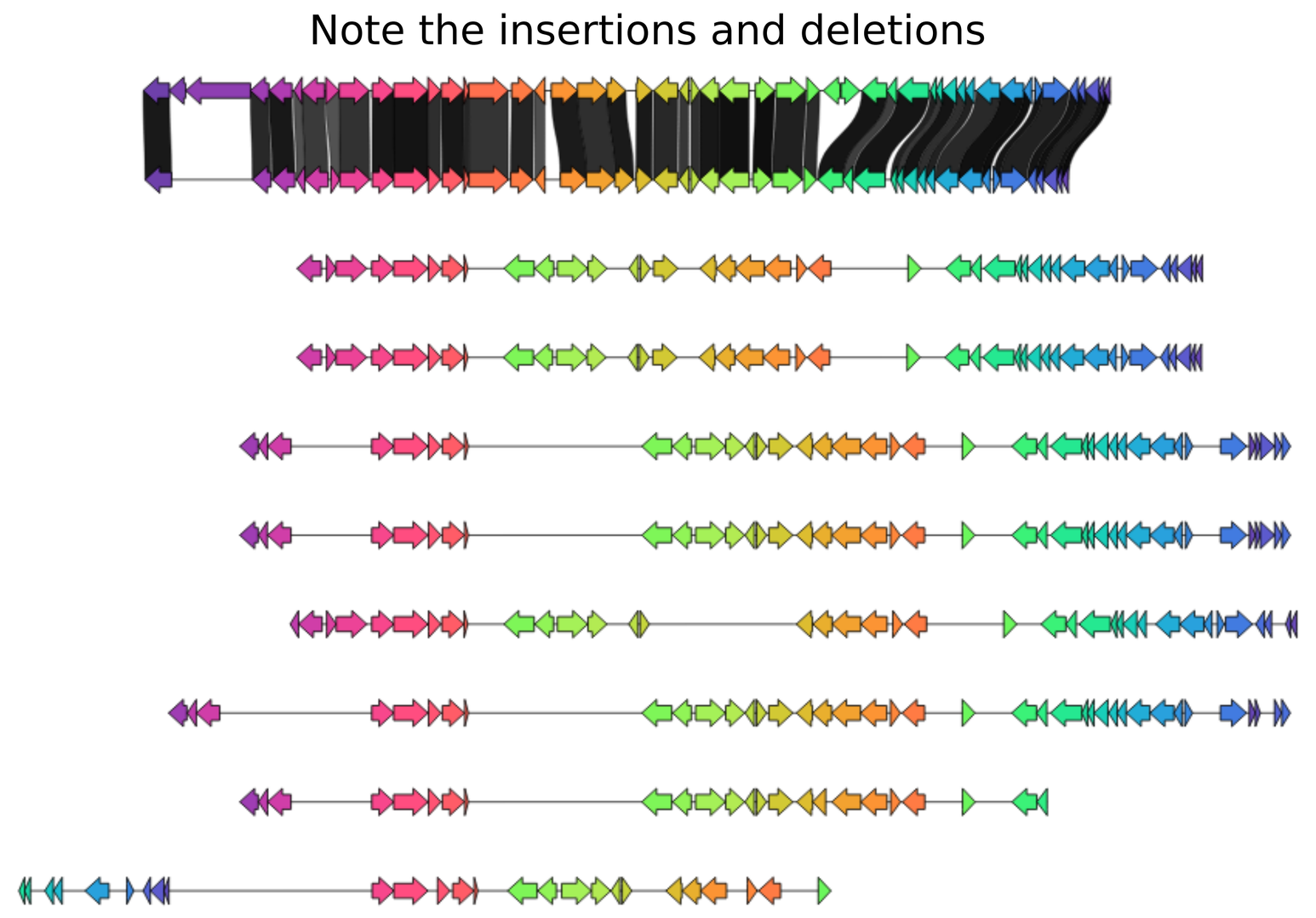
...Neither is BGC "similarity"

The plots below show SocialGene results from searching all MIBiG BGCs against all NCBI genomes associated with MIBiG entries (nearly 300). Within each plot, genes/arrows of the same color represent proteins with high similarity in their HMM annotations. The top line is the MIBiG BGC.

The million dollar question is: what is the relationship between protein similarity, genomic context, and encoded natural product similiarity?

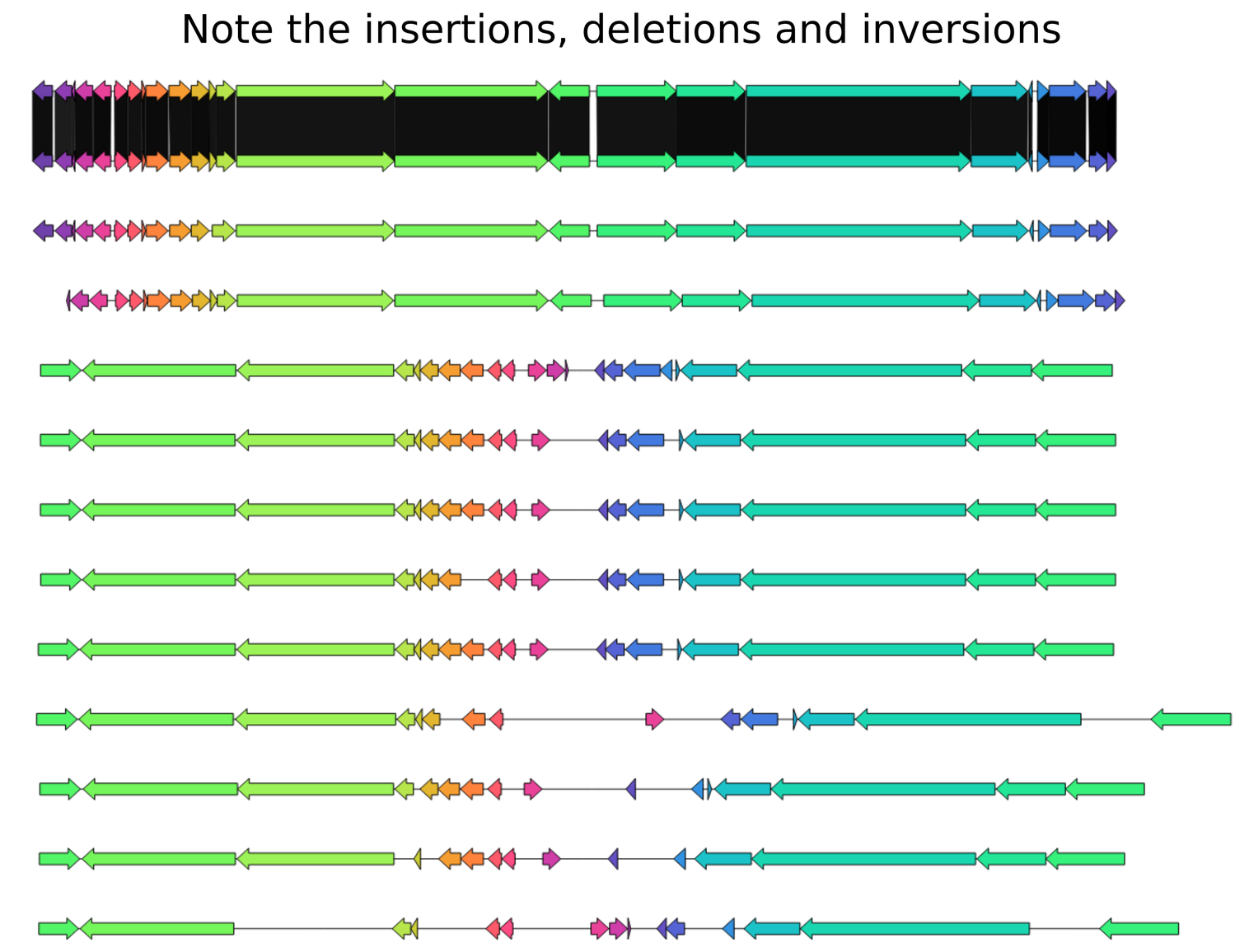
BGC0000279; xantholipin; PKS

Note the insertions and deletions

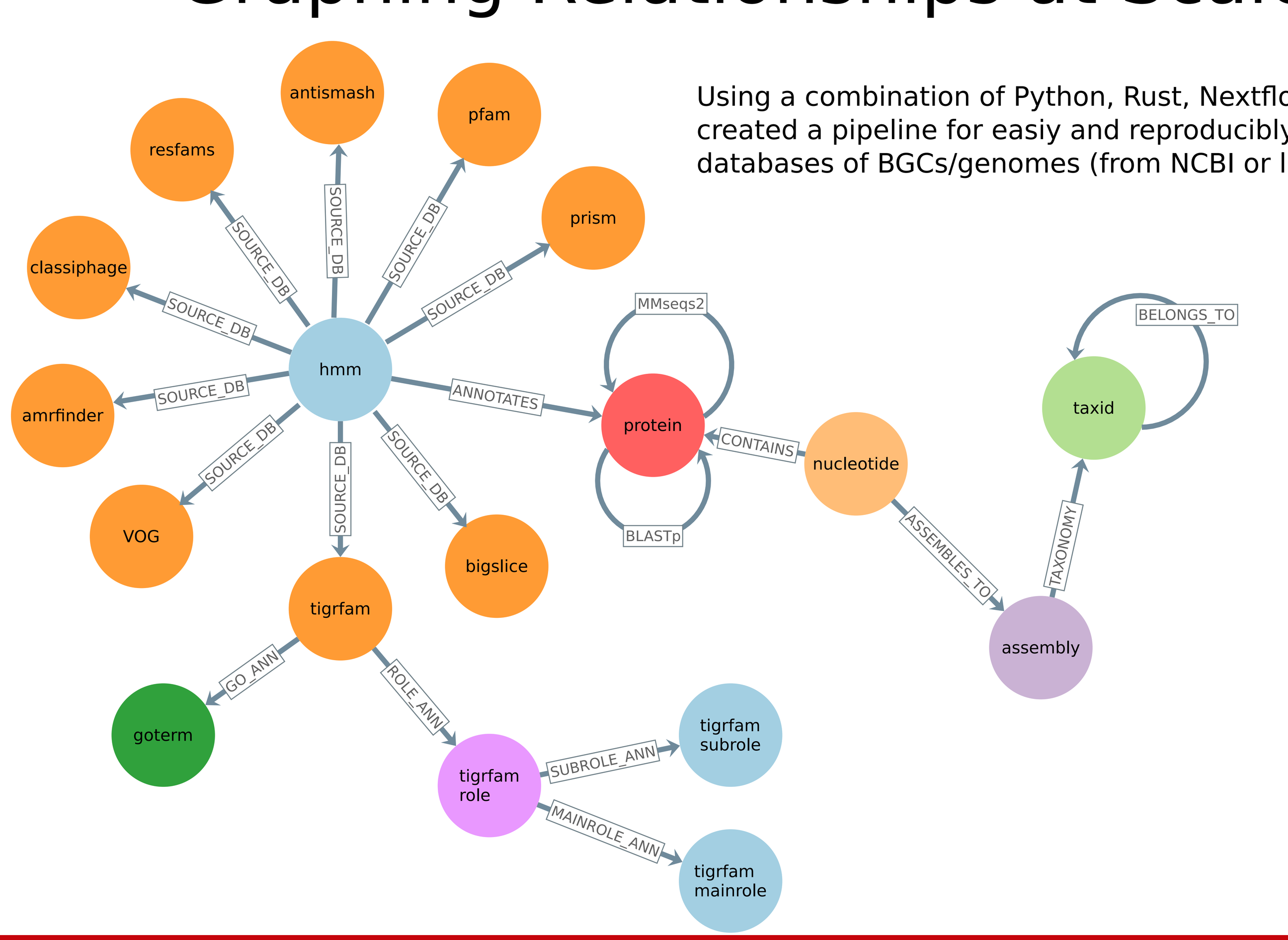


BGC0001770: sceliphrolactam

Note the insertions, deletions and inversions

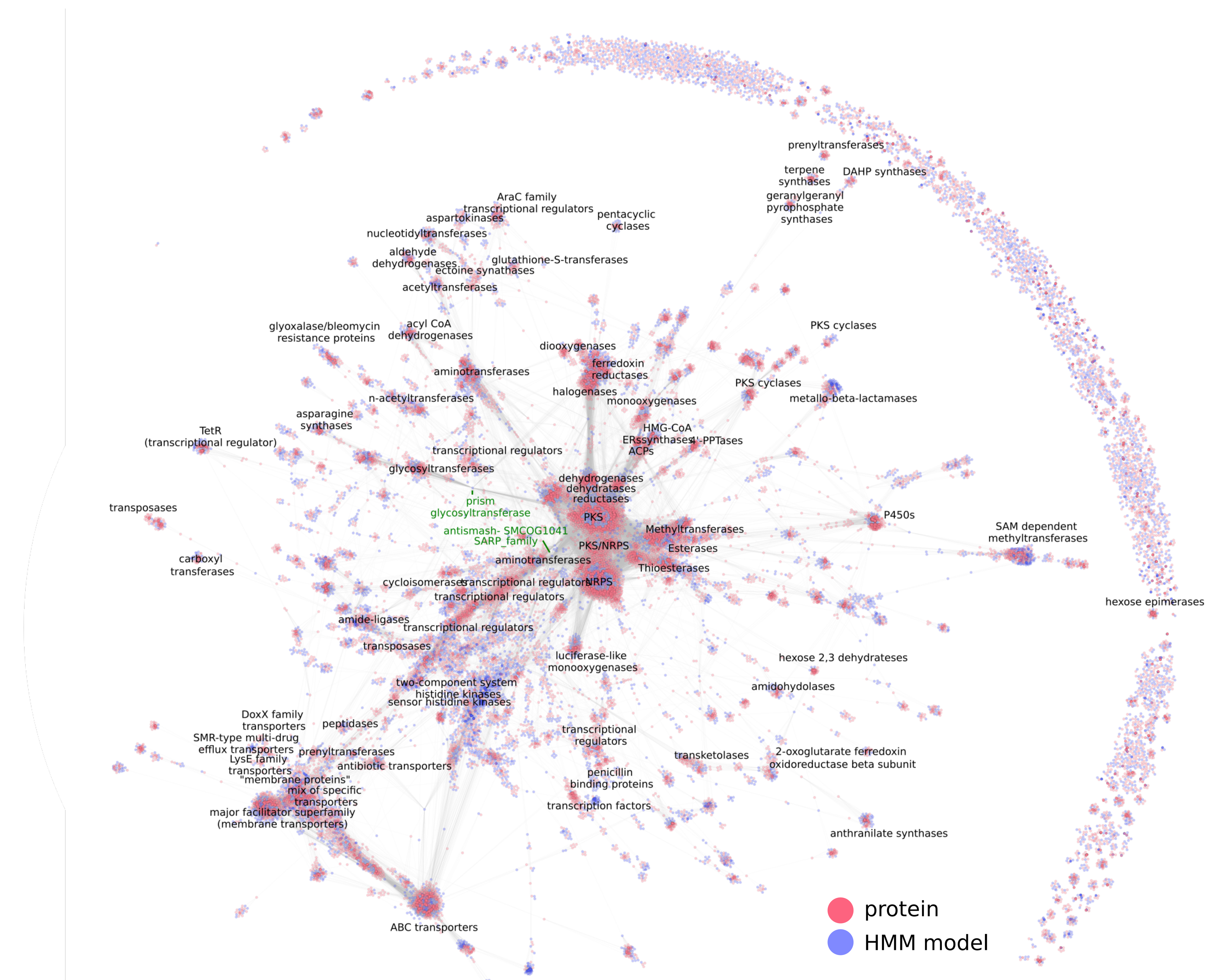


Graphing Relationships at Scale



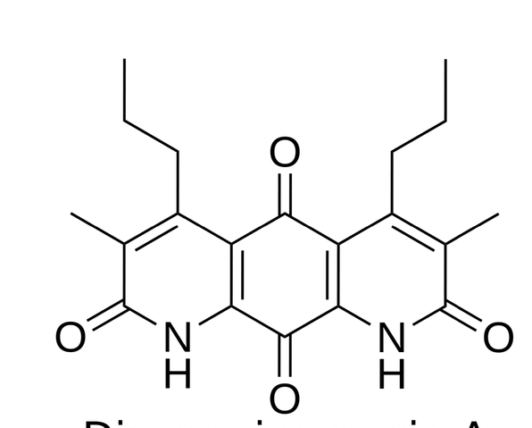
Using a combination of Python, Rust, Nextflow and Neo4j I've created a pipeline for easy and reproducibly creating graph databases of BGCs/genomes (from NCBI or locally provided).

Mapping MIBiG



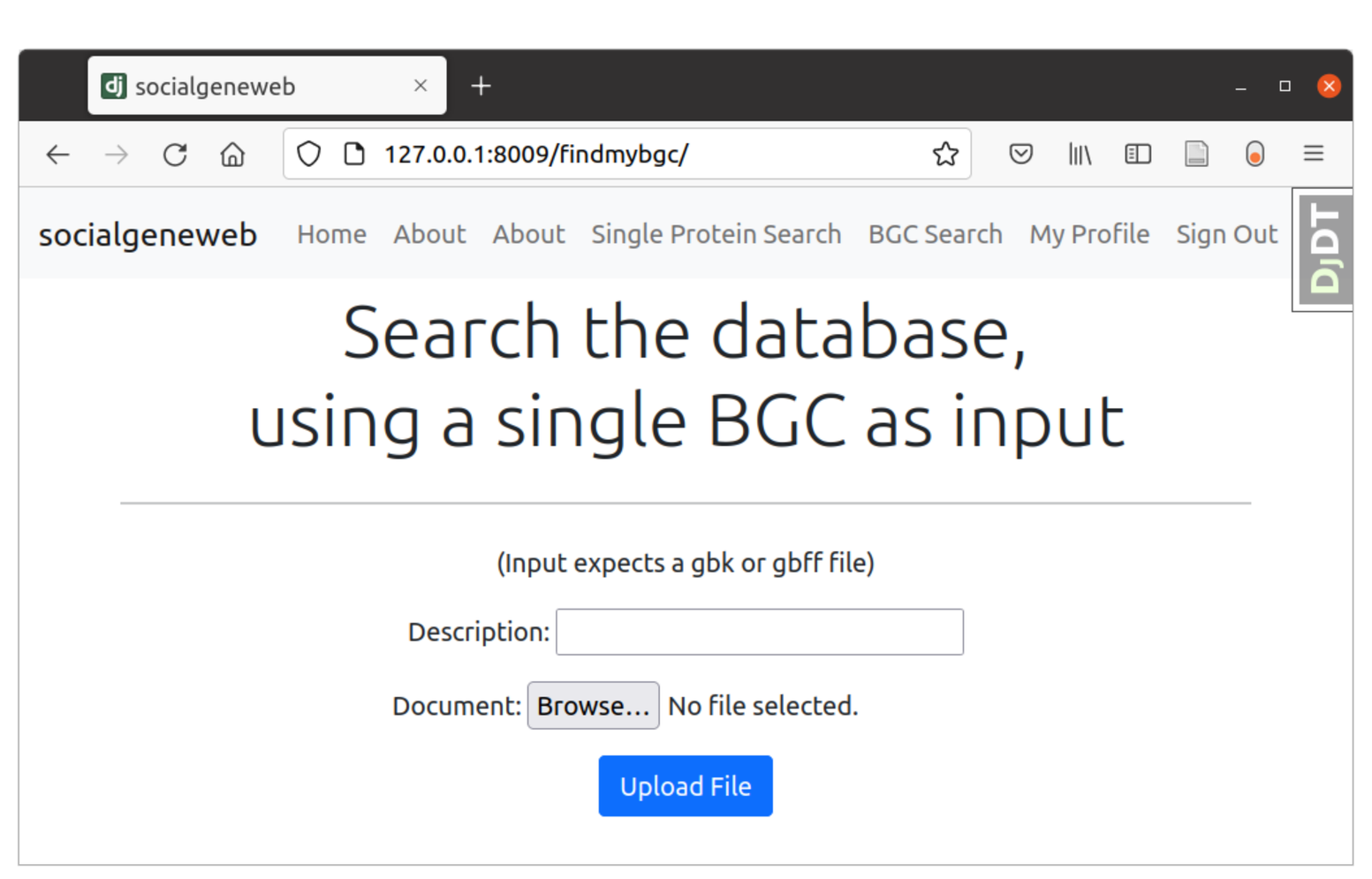
Interactive version:
socialgene.github.io/mibigmap

Targeted BGC search



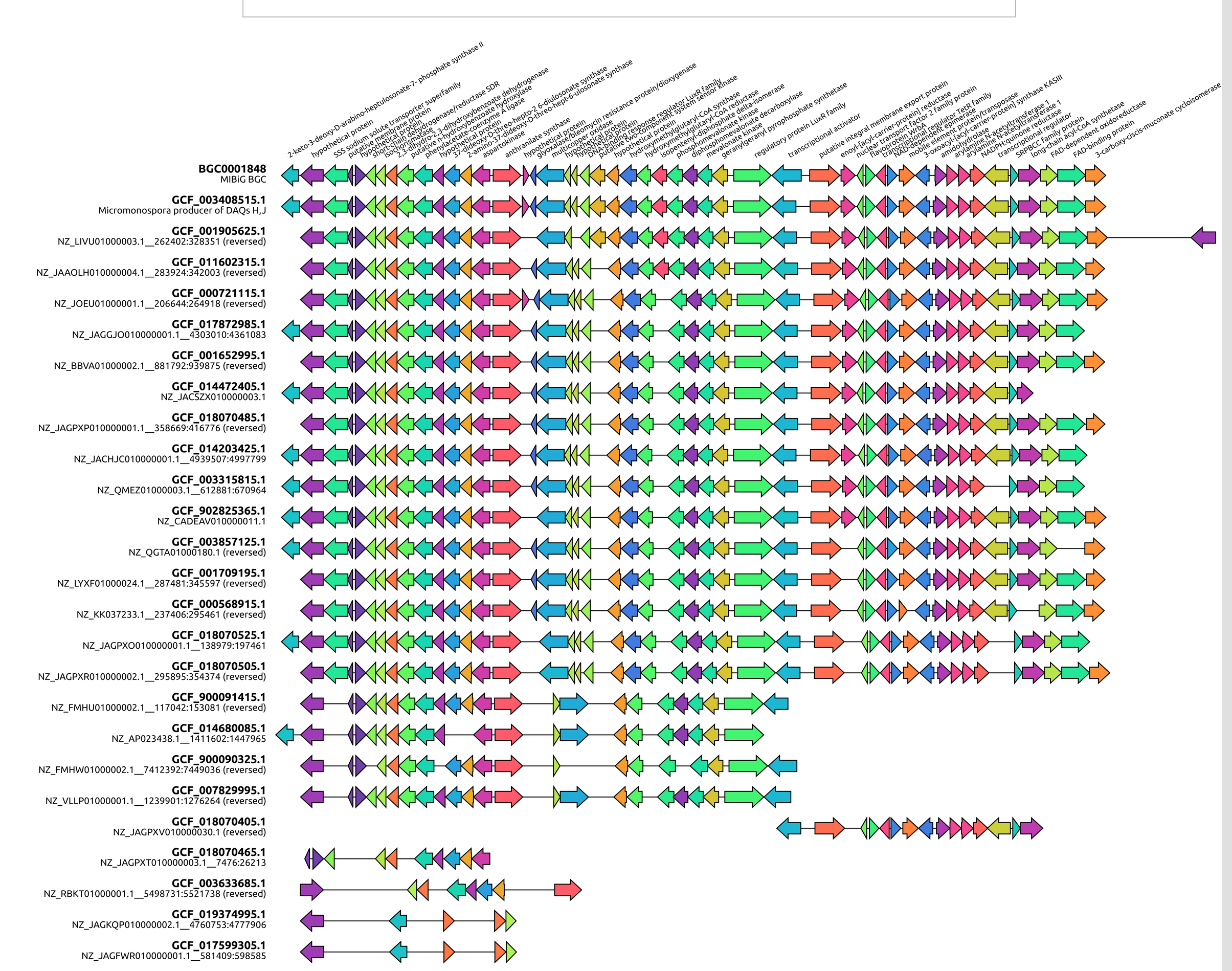
Diazaquinomycin A

	Micromonospora
Genomes	226
Contigs/Scaffolds	34,856
Proteins	1,373,641
Non-Redundant Proteins	1,076,690
HMM Annotations	10,656,754
MMSEQS2	1,009,056
Bidirectional BLASTP (45.3 GB)	420,788,981
Database Creation (Real Duration)	4h 30m
Database Creation (CPU Hours)	133



Diazaquinomycin (DAQ) H and J are selective inhibitors of *Mycobacterium tuberculosis* that stalled in preclinicals due to their horrible aqueous solubility. While just a simple test case, finding BGCs encoding potential chemical analogs could potentially open new routes to compound druggability. DAQs have been found in multiple genera, including *Micromonospora* spp and *Streptomyces* spp.[8-10]

For this test case I uploaded the MIBiG DAQ BGC to Socialgene's web interface, connected to a socialgene database containing all 226 Micromonospora genomes currently available in RefSeq. With its success I then repeated the experiment using >1,200 Streptomyces genomes (all assemblies except contig-quality).

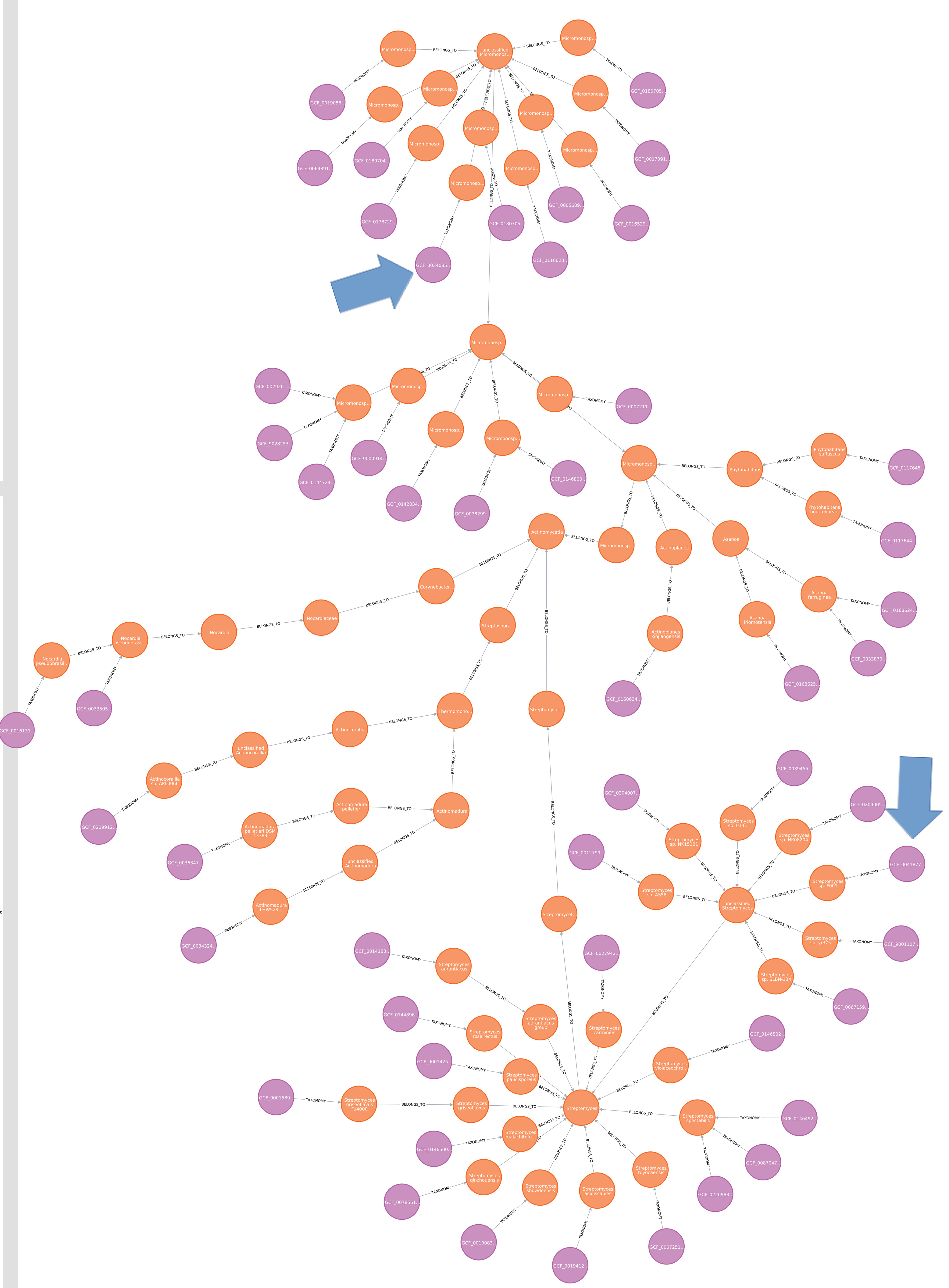


Repository-Scale BGC Search

Annotating a single genome with the PFAM HMM database takes about 20 minutes using a single CPU. With the help of UW-Madison's Center fo High Throughput Computing and the Open Science grid we can annotate hundreds of thousands of genomes in less than a day. Below I show results from searching the same diazaquinomycin BGC as before, except against all genomes from NCBI's RefSeq. Orange nodes represent taxonomic IDs (at different levels) and purple nodes represent genomes that contain a putatively-related BGC.

Center for High Throughput Computing — Open Science Grid

	RefSeq
Genomes	266,668
Contigs/Scaffolds	23,941,594
Proteins	188,429,555
HMM models	25,648
HMM annotations	1,403,423,051
MMseqs2 (bug in MMseqs2)	188,327,165
Contigs to Proteins	971,298,319
Species	49,902
Genera	6,460



CMC is supported by a T15 NLM training grant to the Computation and Informatics in Biology and Medicine Training Program (NLM 5T15LM007359) and R35GM133776