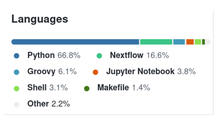
Towards near-instant, repository-scale searching for homologous BGCs with SocialGene

Chase M. Clark, Jason C. Kwan
Division of Pharmaceutical Sciences, School of Pharmacy
University of Wisconsin–Madison,
777 Highland Avenue, Madison, WI 53705, USA

A common workflow in many genomic-based discovery groups is to take a known biosynthetic gene cluster (BGC) and search its individual proteins against NCBI's non-redundant databse (nr) using nucleotide (BLASTn) or protein based alignments (BLASTp). A number of tools have been developed to aid this process (e.g. MultiGeneBlast[1], cblaster[2], clusterfinder[3] etc.) However, when trying to include the genomic context of results when searching large databases (RefSeq >230 thousand genomes; GenBank >1.1 million genomes) the results returned from a single protein search often exceed what these programs and/or the NCBI BLAST API can handle (most searches silently limit the number of results).
Socialgene was built to provide a fast and easy way to search large numbers of genomes for similar individual proteins, and BGCs. It is alignment free, relying instead searching through pre-annotated protein domains via a graph database (other tools using domain similarity include micropan, RAMPAGE, ENTS, CO-ED, ClusterFinder, etc.).[3-5] Additionally, there are no inbuilt assumptions of what a BGC should look like, which allows the flexibility of finding BGCs that are broken/split due to biology or incomplete sequencing/ assembly. Two common aligment-based tools, BLAST (through Diamond[6]) and MMSeqs2[7], are also incorporated and future searching may take advantage of pre-clustering by MMseqs2.

27,580 HMM models — Pfam, TIGRFAM, AMRFinder, PRISM, antiSMASH, BiG-SLiCE, and more?



Mapping MIBiG

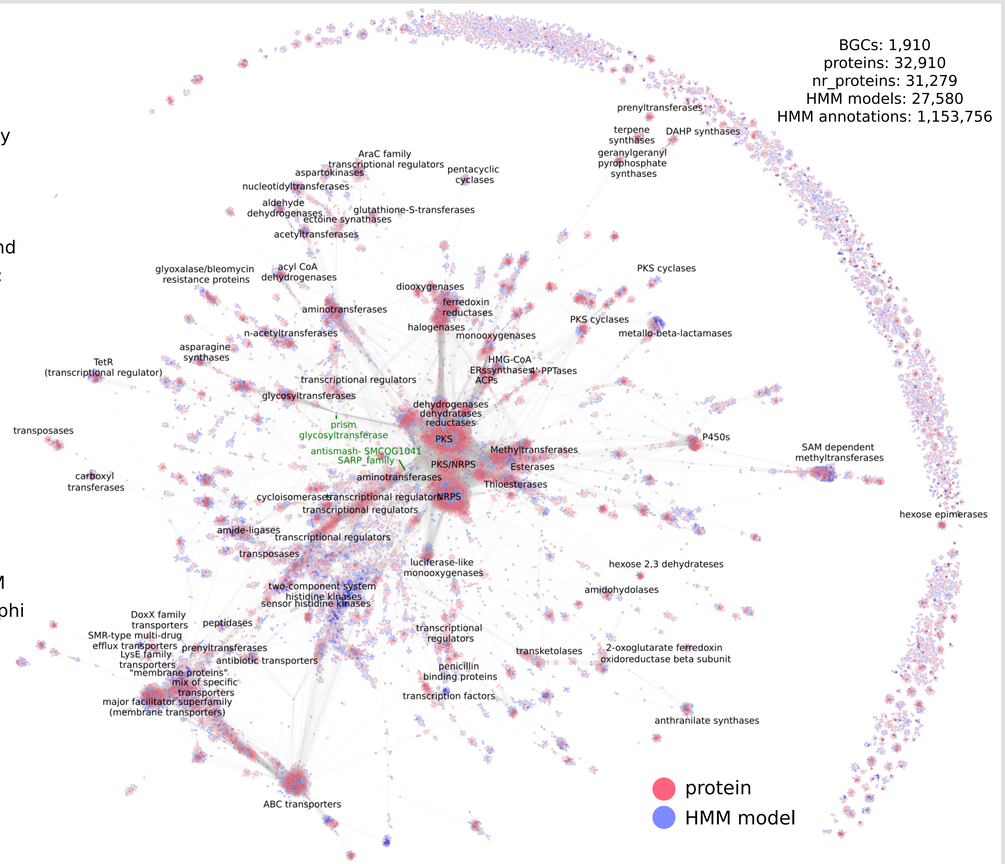
As we're ultimately interested in BGCs, it seemed like a good sanity check would be to see how a SocialGene database "looked" when created with only known BGCs.
Therefore, 1,910 BGCs were downloaded from the Minimum Information about a Biosynthetic Gene cluster (MIBiG) database and the graph database created using SocialGene's Nextflow workflow:

```
nextflow run nextflow \
  --gbk_input 'mibig_gbk_2.0.tar.gz' \
  -profile simple_custom_input,conda
```

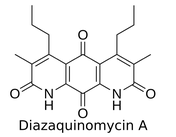
Duration: 22m
CPU Hours: 6.2

The resulting database was queried and all proteins and their HMM annotations exported as a graph/network for visualization with Gephi (the graph to the right). Red circles represent the non-redundant proteins present in MIBiG BGCs, while blue circles repesent HMM models (protein domains/motifs). Resulting clusters were fairly homogenous in protein function and some have been labeled.

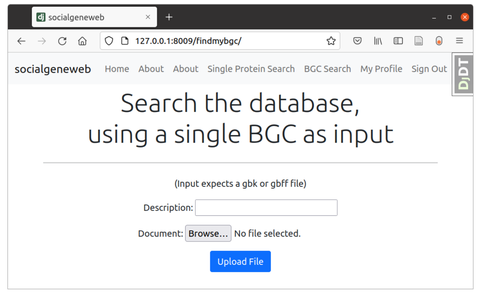
Interactive version:
socialgene.github.io/mibigmap



Searching for Diazaquinomycin BGCs

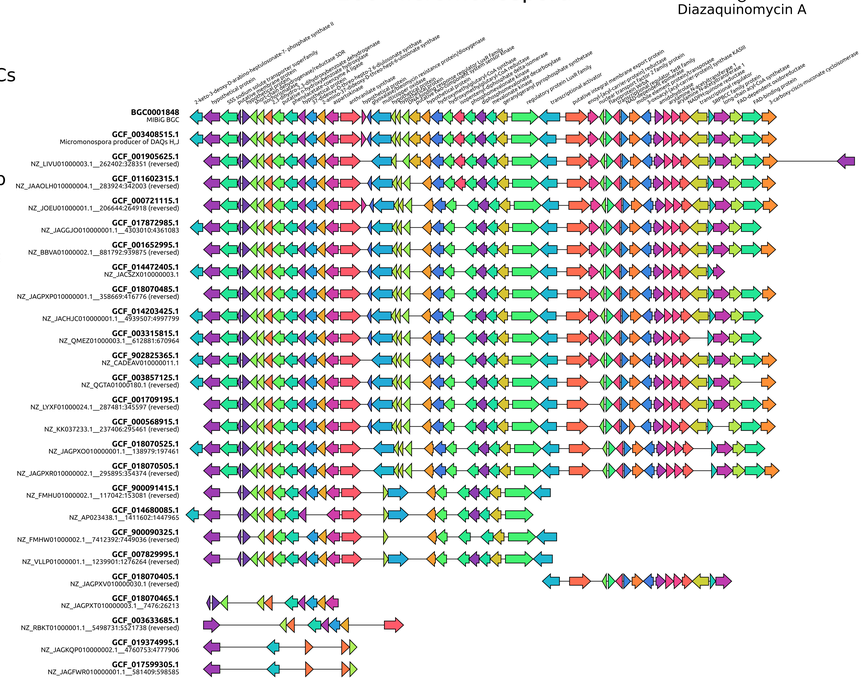


Diazaquinomycin (DAQ) H and J are selective inhibitors of *Mycobacterium tuberculosis* that stalled in preclinicals due to their horrible aqueous solubility. While just a simple test case, finding BGCs encoding potential chemical analogs could potentially open new routes to compound druggability. DAQs 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).

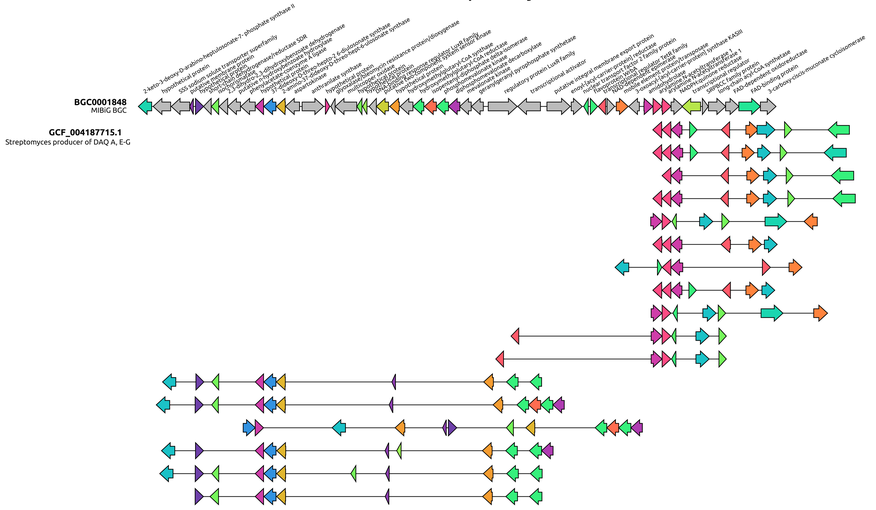


	Micromonospora	*Streptomyces*
Genomes	226	1,202
Contigs/Scaffolds	34,856	186,574
Proteins	1,373,641	8,561,870
Non-Redundant Proteins	1,076,690	6,236,718
HMM Annotations	10,656,754	65,467,422
MMSEQS2	1,009,056	6,234,349
Bidirectional BLASTP (45.3 GB)	420,788,981	Not performed
Database Creation (Real Duration)	4h 30m	22h
Database Creation (CPU Hours)	133	524

Results from a search of >200 Micromonospora



Results from a search of >1200 Streptomyces



References

1. Medema, M. H.; Takano, E.; Breitling, R. Detecting Sequence Homology at the Gene Cluster Level with MultiGeneBlast. Mol. Biol. Evol. 2013, 30 (5), 1218–1223.
2. Gilchrist, C. L. M.; Booth, T. J.; van Wersch, B.; van Grieken, L.; Medema, M. H.; Chooi, Y.-H. Cblaster: A Remote Search Tool for Rapid Identification and Visualization of Homologous Gene Clusters. Bioinformatics Advances 2021, 1 (1).
3. Cimermancic, P.; Medema, M. H.; Claesen, J.; Kurita, K.; Wieland Brown, L. C.; Mavrommatis, K.; Pati, A.; Godfrey, P. A.; Koehrsen, M.; Clardy, J.; Birren, B. W.; Takano, E.; Sali, A.; Linington, R. G.; Fischbach, M. A. Insights into Secondary Metabolism from a Global Analysis of Prokaryotic Biosynthetic Gene Clusters. Cell 2014, 158 (2), 412–421.
4. Scaiewicz, A.; Levitt, M. The Language of the Protein Universe. Curr. Opin. Genet. Dev. 2015, 35, 50–56.
5. de Rond, T.; Asay, J. E.; Moore, B. S. Co-Occurrence of Enzyme Domains Guides the Discovery of an Oxazolone Synthetase. Nat. Chem. Biol. 2021, 17 (7), 794–799.
6. Buchfink, B.; Reuter, K.; Drost, H.-G. Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND. Nat. Methods 2021, 18 (4), 366–368.
7. Steinegger, M.; Söding, J. MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets. Nat. Biotechnol. 2017, 35 (11), 1026–1028.
8. Braesel, J.; Clark, C. M.; Kunstman, K. J.; Green, S. J.; Maienschein-Cline, M.; Murphy, B. T.; Eustáquio, A. S. Genome Sequence of Marine-Derived *Streptomyces* Sp. Strain F001, a Producer of Akashin A and Diazaquinomycins. Microbiol Resour Announc 2019, 8 (19).
9. Braesel, J.; Tran, T. A.; Eustáquio, A. S. Heterologous Expression of the Diazaquinomycin Biosynthetic Gene Cluster. J. Ind. Microbiol. Biotechnol. 2019, 46 (9-10), 1359–1364.
10. Braesel, J.; Crnkovic, C. M.; Kunstman, K. J.; Green, S. J.; Maienschein-Cline, M.; Orjala, J.; Murphy, B. T.; Eustáquio, A. S. Complete Genome of *Micromonospora* Sp. Strain B006 Reveals Biosynthetic Potential of a Lake Michigan Actinomycete. J. Nat. Prod. 2018, 81 (9), 2057–2068.
11. Mullowney, M. W.; Hwang, C. H.; Newsome, A. G.; Wei, X.; Tanouye, U.; Wan, B.; Carlson, S.; Barranis, N. J.; hAinmhire, E.; Chen, W.-L.; Krishnamoorthy, K.; White, J.; Blair, R.; Lee, H.; Burdette, J. E.; Rathod, P. K.; Parish, T.; Cho, S.; Franzblau, S. G.; Murphy, B. T. Diaza-Anthracene Antibiotics from a Freshwater-Derived Actinomycete with Selective Antibacterial Activity toward *Mycobacterium Tuberculosis*. ACS Infect Dis 2015, 1 (4), 168–174.
12. Braesel, J.; Lee, J.-H.; Arnould, B.; Murphy, B. T.; Eustáquio, A. S. Diazaquinomycin Biosynthetic Gene Clusters from Marine and Freshwater Actinomycetes. J. Nat. Prod. 2019, 82 (4), 937–946.
13. Gilchrist, C. L. M.; Chooi, Y.-H. Clinker & Clustermap.js: Automatic Generation of Gene Cluster Comparison Figures. Bioinformatics 2021.

Funding

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