DOE Joint Genome Institute FY15 Q4 Progress Report

Q4: Report on a new computational method for improving the interpretation of microbial, metagenomic, or plant genomes.

Background

Secondary metabolites (SMs) are naturally occurring compounds that have a broad activity spectrum. Certain classes of SMs are good candidates for biofuel production. Others may serve as important agents of communication in symbiotic relationships and interactions between biofuel plant feedstocks and their microbial communities, and as such may be useful tools in the manipulation of these systems. SMs produced by plant-associated bacteria may also be used for biocontrol, thus providing a cost-effective way for controlling plant pathogens and increasing crop yield. Therefore, advances in SM discovery would benefit not only the development of novel biotechnologies, but also contribute to our understanding of complex environments and the networks of communication within them. Traditionally, SMs have been isolated from the plants or the cultured microbes that produce them and/or identified by large-scale analytical screening. However, for microbial SMs this approach has major limitations since up to 99% of microbial organisms found in environmental samples are uncultivable with current methods. Additionally, even for those organisms that can be grown in a controlled environment, their full biosynthetic potential might not be reached due to the absence of unknown environmental conditions or stimuli.

In the last few years, the widespread adoption of high-throughput sequencing technologies has led to an explosion of genomic data of both isolate organisms and microbial communities (metagenomes). The availability of large-scale genomic data, in conjunction with the development of tools to computationally identify and classify biosynthetic gene clusters (BCs) responsible for secondary metabolite production presents a new opportunity for SM discovery. BCs identified using this approach can be cloned or synthetically reconstructed and expressed in heterologous systems, which can then be monitored for the production of the potentially novel SM. In response to a renewed interest in the search for novel microbial SMs, a few databases have been developed. Although some of these efforts provide useful high-quality manually curated annotations, these annotations come at the cost of narrow specificity and often have no or limited maintenance. Most of these systems are limited either to specific classes of organisms (e.g. Streptomyces) or biosynthetic cluster types (e.g. NRPS/PKS). Additionally, some have relatively few records and do not provide the tools for in-depth sequence analyses.

Progress

Here, we outline the progress made in the last year in developing the pipelines for the annotation of microbial genomes with secondary metabolism-related attributes. Furthermore, we have created a data mart within the Integrated Microbial Genomes (IMG) resource, the Atlas of Biosynthetic gene Clusters (IMG-ABC) (Hadjithomas et al., 2015), which contains a set of

databases and toolkits needed for the utilization of these genomic annotations. IMG-ABC can be accessed directly (https://img.jgi.doe.gov/abc) or through its seamless integration with the user-interface structure of the IMG system (http://img.jgi.doe.gov/) via the “ABC” Data Mart. The synergies gained with the IMG integration add an extraordinary value to the IMG-ABC database because the user has immediate access not only to both computationally predicted and experimentally validated BCs and associated SMs, but also to a vast array of integrated functionally and phylogenetically annotated genomic/metagenomic data, gene expression data, and functional genomics data. Users also have access to search-based and statistics-based entry points into both BC and SM objects within IMG. Most analysis tools in IMG are accessible to workflows employing an SM or a BC as a starting point. An important added benefit stemming from IMG-ABC’s integration with IMG is that it will help expose secondary metabolism to the 10,000+ registered IMG users, some of whom may have never before considered studying BCs in their favorite organisms, and will empower them to quickly and easily ask questions using a familiar interface**.**

1. Retrieval of data describing experimentally verified biosynthetic gene clusters and secondary metabolites.

An important aspect of any computational resource with predicted annotations is the existence of a high confidence dataset to be used for comparison. We set out to gather as much experimentally acquired information pertaining to BCs and SMs as possible. We achieved that by developing an automated process for the acquisition of Genbank records that describe biosynthetic gene clusters. We then retrieved the potential names of the produced SMs and used them to query the Pubchem Compound database, in order to retrieve chemical structure and composition description and other relevant information. Additionally, we performed data mining from public databases of biosynthetic gene clusters. A schematic overview of the IMG-ABC object structure and associated attributes for BCs and SMs is illustrated in Figure 1. Lastly, JGI has participated in a consortium to establish a set of standards for the systematic annotation and description of experimentally studied biosynthetic gene clusters and the secondary metabolites they produce (Medema et al, 2015).

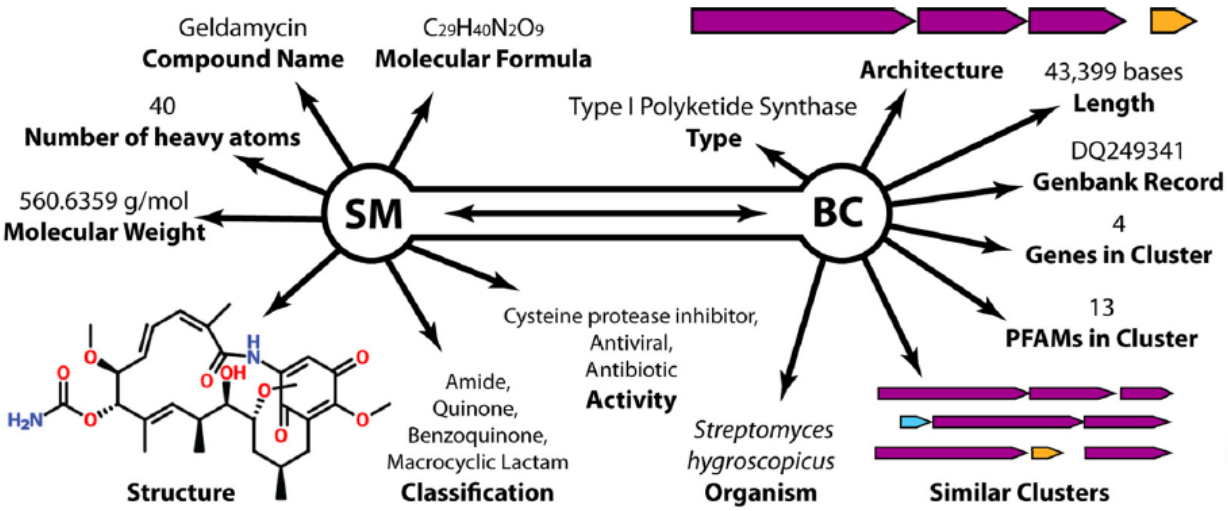

Figure 1 IMG-ABC object structure overview. IMG-ABC contains two main classes of objects with a variety of attributes describing predicted and experimentally studied biosynthetic gene clusters and the secondary metabolites associated with the latter.

2. Computational prediction of putative secondary metabolite biosynthetic gene clusters in genomes and metagenomes.

In the last few years, novel and improved algorithms for the prediction and annotation of biosynthetic gene clusters in genomic sequences have been developed. However, none of these tools has an optimal combination of accuracy and scalability to predict biosynthetic clusters in thousands of genomes in the IMG database. Therefore we used the highly scalable ClusterFinder (Cimermančič et al, 2014) to provide initial BC predictions for all isolate genomes and metagenomes in IMG and IMG/M, respectively. ClusterFinder predictions are based on the Pfam annotations available for all genomes within the integrated context of IMG. In order to improve the accuracy of ClusterFinder predictions, BCs that contained fewer than 6 genes or had a prediction probability of less than 0.3 were removed. Additionally, BC content was analyzed to identify Pfam categories that are not known to be associated with secondary metabolism but are present as positive training features within ClusterFinder. These included Pfam categories such as prophage proteins (PF04883, PF05709, and PF06199), protein secretion systems (PF08817, PF10140, and PF10661), inorganic ion transport proteins (PF02421 and PF11604), and families representing DNA/RNA polymerases, whose presence leads to false-positive BC predictions. For isolate genomes this refined set of predicted BCs generated by ClusterFinder has been classified by the type of biosynthetic enzymes using the highly accurate, but poorly scalable AntiSMASH tool (Blin et al, 2013).

3. IMG-ABC content.

Recent advances in sequencing technologies led to a rapid increase in the number of sequenced microbial genomes and metagenomes. Mining these sequences for novel biosynthetic enzymes and BCs represents a significant challenge for researchers. However, it can be greatly facilitated if computational predictions of candidate genes and gene clusters in

their genomes of interest are provided in combination with the tools for comparative analysis and manual exploration. To achieve this goal we have undertaken a computationally intensive BC annotation effort in order to establish of a large database of predicted and experimentally verified BCs. This database, called IMG-ABC, contains more than 700,000 BCs in publicly available records (Table 1), most of them in bacterial isolate genomes from 57 phyla reflecting the diversity of available genomic sequences in the IMG database. The integration with the GOLD database allows linking the secondary metabolism annotations with metadata associated with genomes and metagenomes. A subset of BCs from isolate genomes has been assigned to one or more BC enzymatic types based on the presence of signature core enzymes. An integration with IMG data and extensive analysis tools provides multiple useful navigation capabilities and allows researchers to zoom in on BCs found in their genomes and metagenomes of interest, review specific BC classes, identify similar BCs in other genomes, analyze genes and enzymes, as well as identify BCs that are likely to produce metabolites similar to their compounds of interest. Some of the unique capabilities implemented in IMG-ABC are described below.

Domain (# public samples with BCs)	Predicted BCs	Experimental BCs
Metagenomes (2174)	251,075	-
Archaea (485)	4,122	1
Bacteria (20372)	417,514	204
Eukaryota (181)	41,880	4
Archaeal Genome Fragments (1)	-	1
Bacterial Genome Fragments (1111)	389	1,076
Eukaryotic Genome Fragments (140)	29	138
Archaeal Plasmids (0)	-	-
Bacterial Plasmids (195)	269	-
Eukaryotic Plasmids (7)	7	-
Other Plasmids (1)	1	-
Viruses (838)	1,222	-
Total	**716,508**	**1,424**

Table 1 Distribution of experimentally validated and predicted Biosynthetic Gene Cluster annotations within various domains of life, genome fragments, and plasmids.

4. IMG-ABC search functions.

The following search functions allow users to focus their work on a subset of the annotations:

a. **Searching biosynthetic clusters and secondary metabolites by annotation.**
The IMG-ABC database contains a large number of experimentally validated and predicted BCs with the former being connected to the SMs that they are known to

produce. To enable users to efficiently narrow their search based on their specific interests, two search interfaces have been implemented. Both SMs and BCs can be searched by a combination of chemical names, SM chemical classification, SM activity, number of atoms, molecular weight and chemical formula, while the search space can be limited to specific target clades through a phylogenetic tree-based menu. Additionally, BCs are searchable by genomic features such as the number of genes, length of the BC, probability of prediction, and type of enzymatic mechanism in the BC assigned by the AntiSMASH tool. Since there can be multiple combinations of BC types assigned to a cluster, we also provide users with the option to perform exact or inexact ("AND"/"OR") searches for BCs with hybrid types. Another useful option is the ability to limit the results based on the presence of a single Pfam or combinations thereof, which enables "fishing" for enzymatic activities that may not be included in the AntiSMASH classification routine and therefore, can aid in the discovery of new enzymatic mechanisms and/or novel chemical structures. The results are displayed in a tabular form and, in the case of BC searches, users can add selected BCs to the "Scaffold Cart" or visualize their gene neighborhoods.

b. **Searching IMG-ABC using a chemical structure.** The ability to use a chemical structure to query a genomic database is another new and exciting feature introduced by IMG-ABC. A known SMILES descriptor (provided by the user or retrieved through an external interface such as NCBI PubChem's Sketcher) is used to search compounds based on structural similarity computed using the Tanimoto similarity score to produce an interactive listing of similar SMs and related BCs sorted by the similarity score. This search is powered through an implementation of ChemMineR functions and facilitated by the pre-computed atom pair descriptors for all compounds in the IMG-ABC database. SMILES strings entered by users are converted to structure-data file format (SDF) and then to atom pair descriptors which are used to search either all secondary metabolites, or all IMG compounds, depending on the user's input.

Conclusion

Annotation of biosynthetic gene clusters in newly loaded genomes and metagenomes and their maintenance is now part of the regular loading and maintenance schedules of IMG. The growing number of predicted BCs, in conjunction with continuous development of the analysis and search functions available through the system, will ensure that IMG-ABC will always have the latest and most complete publicly available information for studies of secondary metabolism in microbial genomes and metagenomes.

DOE JGI team

Michalis Hadjithomas, Ken Chu, Victor Markowitz, Natalia Ivanova, Nikos Kyrpides

References

(in bold JGI Authors)

Hadjithomas M, Chen IM, Chu K, Ratner A, Palaniappan K, Szeto E, Huang J, Reddy TB, Cimermančič P, Fischbach MA, **Ivanova NN, Markowitz VM, Kyrpides NC, Pati A.** (2015) IMG-ABC: A Knowledge Base To Fuel Discovery of Biosynthetic Gene Clusters and Novel Secondary Metabolites. MBio. 2015 Jul 14;6(4):e00932. doi: 10.1128/mBio.00932-15.

Medema MH, …, **Hadjithomas M**, …, **Pati A,** …, **Kyrpides NC**, …, Glöckner FO. (2015) Minimum Information about a Biosynthetic Gene cluster. Nature Chem Biol. 2015 Aug 18;11(9):625-31. doi: 10.1038/nchembio.1890

Cimermančič P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, **Mavrommatis K,Pati A,** Godfrey PA, Koehrsen M, Clardy J, Birren BW, Takano E, Sali A, Linington RG,Fischbach MA. (2014) Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell. 2014 Jul 17;158(2):412-21. doi: 10.1016/j.cell.2014.06.034.

Blin, K., Medema, M. H., Kazempour, D., Fischbach, M. A., Breitling, R., Takano, E., & Weber, T. (2013). antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Research 2013; 41(Web Server issue), W204–W212. doi: 10.1093/nar/gkt449.