Journal of Industrial Microbiology & Biotechnology
https://doi.org/10.1007/s10295-017-1993-1

SIMB
Society for Industrial Microbiology and Biotechnology

NATURAL PRODUCTS

Output ordering and prioritisation system (OOPS): ranking biosynthetic gene clusters to enhance bioactive metabolite discovery

Alejandro Peña[1] · Francesco Del Carratore[1] · Matthew Cummings[1] · Eriko Takano[1] · Rainer Breitling[1]

Received: 17 October 2017 / Accepted: 26 November 2017

Abstract

The rapid increase of publicly available microbial genome sequences has highlighted the presence of hundreds of thousands of biosynthetic gene clusters (BGCs) encoding valuable secondary metabolites. The experimental characterization of new BGCs is extremely laborious and struggles to keep pace with the in silico identification of potential BGCs. Therefore, the prioritisation of promising candidates among computationally predicted BGCs represents a pressing need. Here, we propose an output ordering and prioritisation system (OOPS) which helps sorting identified BGCs by a wide variety of custom-weighted biological and biochemical criteria in a flexible and user-friendly interface. OOPS facilitates a judicious prioritisation of BGCs using G+C content, coding sequence length, gene number, cluster self-similarity and codon bias parameters, as well as enabling the user to rank BGCs based upon BGC type, novelty, and taxonomic distribution. Effective prioritisation of BGCs will help to reduce experimental attrition rates and improve the breadth of bioactive metabolites characterized.

Keywords Genomics · BGC · Synthetic biology · Natural products · Prioritisation

Introduction

Bioactive secondary metabolites are the major source for a diverse range of drugs. The growing wealth of genome sequence data revealed an unexpected diversity of biosynthetic gene clusters (BGCs) potentially responsible for the production of an even larger range of biochemicals [8]. Bacterial and fungal genomes have historically been the source of many medicines; however, most low-hanging fruit has already been picked [1], and classical routes to natural product-based drug discovery generated little in the last two decades [9, 13]. A genome-driven approach, linking biosynthetic gene clusters to specialised metabolites and powered by synthetic biology, might provide a new impetus to natural product discovery [4, 5, 11, 13], enabling access to an unexplored pool of silent, cryptic, and poorly expressed BGCs [14]. Nowadays, BGCs can be systematically predicted from DNA sequence thanks to freely available cluster mining tools (antiSMASH [3], BAGEL [19], CASSIS and SMIPS [18], CLUSEAN [17], ClusterFinder [8], etc.). The widespread use of these tools has provided an unprecedented view of the global distribution of specialised metabolites [8] and is unveiling the complexity of eukaryotic and prokaryotic secondary metabolism. Irrespective of the approach used to identify new BGCs of interest, shortlisting clusters to focus on in subsequent experimental work is a laborious but necessary task. There is no standardised way in which outputs from BGC prediction software can be visualised, filtered and prioritised based on molecular biological principles. Here, we described a freely available and easy to use prioritisation pipeline for biosynthetic gene clusters which uses seven parameters describing the molecular biology of BGCs (number of genes, CDS length, G+C content, codon bias. similarity to known clusters, self-similarity, and phylogenetic diversity) to allow a flexible, user-specific prioritisation of large numbers of BGCs according to properties that are relevant for synthetic biology and secondary metabolite discovery.

Alejandro Peña, Francesco Del Carratore and Matthew Cummings contributed equally to this work.

✉ Rainer Breitling
rainer.breitling@manchester.ac.uk

Francesco Del Carratore
francesco.delcarratore@postgrad.manchester.ac.uk

[1] Manchester Institute of Biotechnology, School of Chemistry, Faculty of Science and Engineering, University of Manchester, Manchester M1 7DN, UK

Published online: 18 December 2017

Materials and methods

The OOPS software is a stand-alone Java-based application built with an embedded web server and accordingly is compatible with all operating systems. OOPS should be considered as an extension of the widely-used antiSMASH pipeline [3], and the input used for OOPS is BGCs obtained with the antiSMASH pipeline.

Prioritisation protocol

After optional filtering of the BGCs according to the cluster type, i.e., the predicted chemical class of the end compound, all seven prioritisation metrics are computed (or extracted from the antiSMASH output) for each BGC. The clusters can be easily ranked according to each metric in ascending or descending order according to the users choice. The final score associated with each BGC is then computed as the simple weighted sum of all the ranks obtained in the previous step. The weighted scoring system provides the user with complete control over which parameters enter the final prioritisation, and to which extent they dominate the results. This allows flexible adjustments of BGC prioritisation according to the specific down-stream analysis envisaged (e.g., some users might search for novel chemistry, independent of the genetic structure of the BGC, while others are looking only for clusters that are easy to manipulate using synthetic biology tools and, therefore, would want to prioritise BGC with few ORFs and relatively short total protein coding sequences). By clicking on any BGC shown in the ranked list by OOPS, the user can also explore the original antiSMASH output related to it, including, for example, links to information about the genomic context and domain architecture. The prioritisation by OOPS assumes that genomes have been correctly assembled, and when interpreting the output, the user should be aware of the possibility of assembly errors affecting the results, especially in the case of type I PKS and NRPS BGCs in unfinished genomes.

Guanine and cytosine content

The G+C content for each cluster is computed from the antiSMASH output simply as the ratio of G+C DNA base pairs over the length of the predicted BGC. If no reference species is selected by the user, BGCs are ranked by their total G+C content. Otherwise, the G+C content of the reference species is obtained from the Kazusa codon usage database (http://www.kazusa.or.jp/codon/), and the clusters are ranked according to the absolute difference between the G+C content in the cluster and the G+C content of the reference species. This allows, e.g., the prioritisation of BGCs that are maximally similar in G+C content to a prospective heterologous host species.

Codon bias

The codon bias parameter can only be used for the prioritisation if a reference species is selected. The codon usage table is computed by OOPS for each BGC, while the table for the species of interest, typically the intended heterologous expression host, is downloaded from the Kazusa website (http://www.kazusa.or.jp/codon/). The clusters are than ranked according the BGC codon bias score computed as follows:

$$\text{BGC codon bias score} = \sum_{i=1}^{64} \mid x_{\text{BGC},i} - x_{\text{species},i} \mid \qquad (1)$$

where $x_{\text{BGC},i}$ and $x_{\text{species},i}$ represent the usage percentage of the ith codon in the cluster and in the species, respectively. Similar to the G+C content, this allows the prioritisation of BGCs that have a codon bias similar to an intended host species, but it could also be employed more creatively, e.g., to prioritise clusters that match the coding patterns of a particular group of organisms with interesting known bioactivities.

Similarity to known cluster

OOPS retrieves from the antiSMASH output the percent similarity (calculated analogously to MultiGeneBlast [12]) of the most similar BGC with a known end product. Then, the final score for this ranking parameter is calculated as the absolute difference between the similarity preference defined by the user (%PS) and the similarity of the most similar cluster (%known cluster) calculated by OOPS:

$$\text{Known cluster score} = \mid \%\text{PS} - \%\text{known cluster} \mid . \qquad (2)$$

This enables users to prioritise BGCs that hit the "sweet spot" of similarity to known clusters, which will differ depending on the specific application scenario: some users will search minor, but important variations of established known compounds; others prefer clusters that are completely different from anything studied before and will have a %PS of zero.

Self-similarity

This parameter is used to prioritise the clusters according to how similar they are to themselves in terms of nucleotide sequences. This metric is computed using a modified version of the Smith–Waterman algorithm [15] to find all suboptimal local alignments of length greater or equal to a user-defined minimum. This parameter allows to eliminate (or prefer)

BGCs with a large number of internal repeats, which might be challenging to engineer genetically (or might be chemically particularly interesting, depending on the use case).

Phylogenetic diversity

OOPS uses the cluster blast output provided for each BGC by antiSMASH [3], which contains all the clusters that are substantially similar to the BGC of interest, to compute a phylogenetic diversity score. The taxonomic identity of the host species of all the associated clusters is obtained from the RESTful services provided by the EMBL–EBI databases [6, 7] and used to build a simplified phylogenetic tree comprising all clusters, using only the taxonomic ranks provided by the database, and collapsing all intermediate nodes. The metric chosen to represent the phylogenetic diversity is the number of nodes of the tree above the species level. A high score indicates that similar clusters are very widespread across the tree of life, while low numbers indicate that similar clusters are found only in a small set of closely related species. Whether this parameter is used in ascending or descending order for the final prioritisation will again strongly depend on the envisaged application. All prioritisation options are accessible via a unified intuitive user interface (Fig. 1).

Availability

OOPS is available at https://github.com/alexcpa/antiSMASH-OOPS. A user guide and tutorial is provided as supplementary material.

Precomputation of actinobacterial BGCs

To illustrate the potential uses of our software, we provide a precomputed data set containing all actinobacterial genomes present within the antiSMASH database [2], consisting of a total of 5903 predicted BGCs. Expression of heterologous actinobacterial BGCs in various *Streptomyces coelicolor* strains [10, 16] is a common first strategy when characterizing new secondary metabolites, and no *S. coelicolor* strain is present in the data set; thus, *S. coelicolor* A3(2) was chosen as our reference species for precomputation of codon bias and G+C content parameters. In addition, to provide good resolution of BGC self-similarity (detection of relevant internal repeats), a

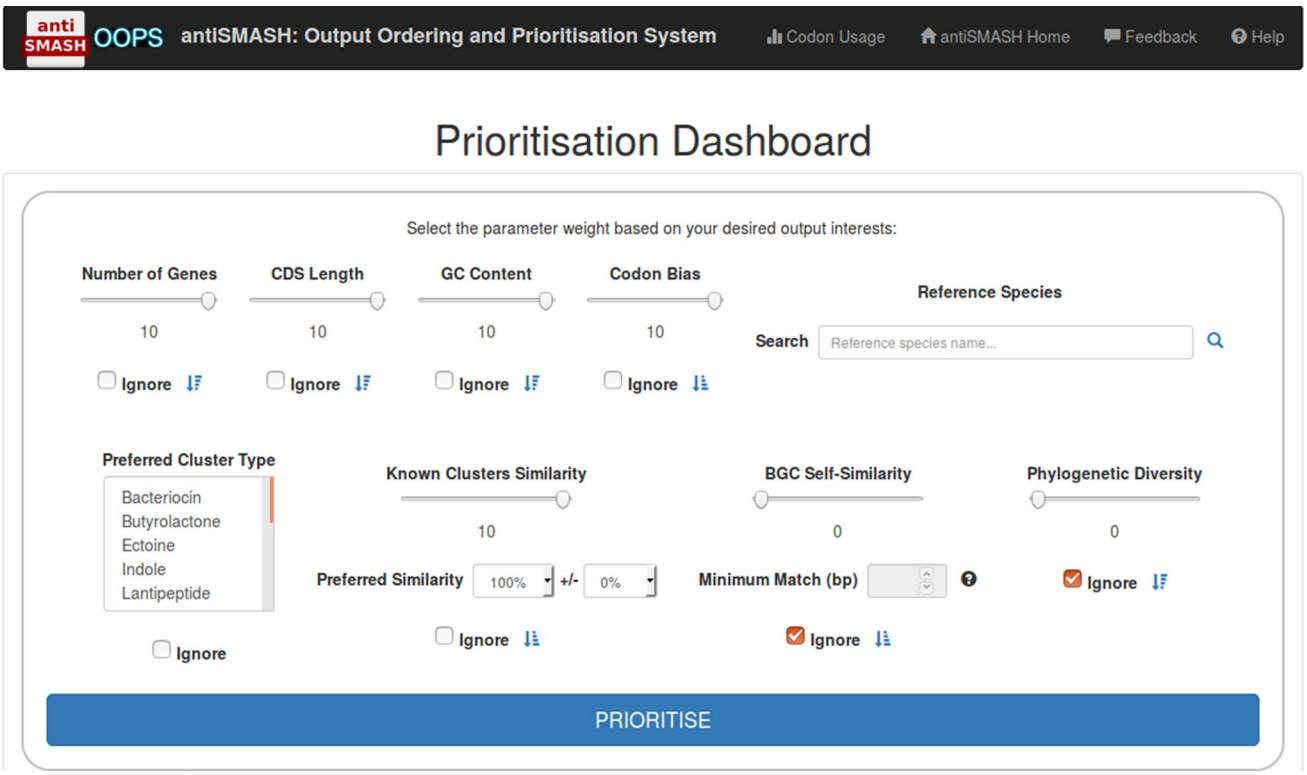

Fig. 1 OOPS graphical user interface. The weighing of each prioritisation parameter can be adjusted using the slide bar, or ignored by checking the associated box. The sorting order (ascending or descending) is specified by clicking the blue arrow icon. The reference species is chosen by typing the species name in the "species" field and is relevant for codon bias and GC content parameters only. Multiple BGC types can be chosen using the "preferred cluster type" field by holding shift and selecting additional BGC classes

threshold of 30 bp was used. Changing the reference species or self-similarity minimum match threshold, when using the large precomputed data set will result in lengthy prioritisation (> 4 days), and therefore, it is not allowed in this case. The precomputed data set here described is available for download at https://doi.org/10.5281/zenodo.1000774. Applying OOPS prioritisation to the precomputed data set allows the user, e.g., to rapidly find out that the two gene clusters with the largest total coding sequence are CP005929.1-C12 from *Actinoplanes* sp. N902-109 (259002 bp) and CP002162.1-C4 from *Micromonospora aurantiaca* (216780 bp; total run-time on a desktop computer: ~ 5 sec). The antiSMASH outputs related to these two clusters can be found at http://antismash-db.secondarymetabolites.org/output/CP005929/index.html#cluster-12 and http://antismash-db.secondarymetabolites.org/output/CP002162/index.html#cluster-4. The user could also find out that the cluster with the largest amount of self-similarity (internal repeats) is CP003899.1-C22 from *Mycobacterium liflandii* (self-similarity score 218870; longest alignment 2000 bp, due to numerous near-identical internal domain duplications in one of its Type I polyketide synthases), which can be found at http://antismash-db.secondarymetabolites.org/output/CP003899/index.html#cluster-22. However, one can also quickly find the answer to more complex user-defined queries that combine a number of criteria in a custom-weighted manner. For instance, if an imaginary user decides that the optimal target for genetic engineering would be a cluster that has a G+C content and codon bias similar to that of *Streptomyces coelicolor* A3(2) (weight 20% each), shows the widest possible phylogenetic distribution (10%), has a minimum amount of internal repeats (20%), and is around 40% similar to the most similar chemically characterized gene cluster (30%), she would within ~ 7 s find out that the closest match to these criteria is CP011492.1-C21 from *Streptomyces* sp. CNQ-509 (found at http://antismash-db.secondarymetabolites.org/output/CP011492/index.html#cluster-21), which has 0.16% difference in G+C compared with *Streptomyces coelicolor* A3(2), a codon bias score of 107.12, a self-similarity score of 0 (i.e., contains no internal repeats above the user-defined threshold), a phylogenetic diversity score of 112 and a known cluster score of 20.36% (i.e., its maximum similarity to a known cluster is 19.64%). At the same time, the user is provided with a long list of next-best matches from which to select potential target clusters, and the opportunity to adjust the prioritisation criteria to fine-tune the selection.

Processing commercially sensitive data

All input data remain offline and are hosted locally by the user, so that commercially sensitive sequences can be analysed and prioritised using the OOPS pipeline.

Acknowledgements AP was funded by CONACyT (Consejo Nacional de Ciencia y Tecnologa), Mexico. MC was funded by a BBSRC DTP studentship (Grant No. BB/J014478/1). This is a contribution from the Manchester Centre for Synthetic Biology of Fine and Speciality Chemicals (SYNBIOCHEM) and acknowledges the Biotechnology and Biological Sciences Research Council (BBSRC) and Engineering and Physical Sciences Research Council (EPSRC) for financial support (Grant No. BB/M017702/1).

References

1. Baltz RH (2006) Marcel Faber roundtable: is our antibiotic pipeline unproductive because of starvation, constipation or lack of inspiration? J Ind Microbiol Biot 33(7):507–513
2. Blin K, Medema MH, Kottmann R, Lee SY, Weber T (2016) The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. Nucleic Acids Res 45(d1):D555–D559
3. Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, Duran S, Hernando G, de los Santos EL, Kim HU et al (2017) antiSMASH 4.0 - improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Res 45(W1):W36–W41
4. Breitling R, Takano E (2015) Synthetic biology advances for pharmaceutical production. Curr Opin in Biotech 35:46–51
5. Breitling R, Takano E (2016) Synthetic biology of natural products. CSH Perspect Biol 8(10):a023994
6. Brooksbank C, Cameron G, Thornton J (2009) The European Bioinformatics Institute's data resources. Nucleic Acids Res 38((suppl–1)):D17–D25
7. Chojnacki S, Cowley A, Lee J, Foix A, Lopez R (2017) Programmatic access to bioinformatics tools from EMBL-EBI update: 2017. Nucleic Acids Res 45(W1):W550–W553
8. Cimermancic P, Medema MH, Claesen J, Kurita K, Brown LCW, Mavrommatis K, Pati A, Godfrey PA, Koehrsen M, Clardy J et al (2014) Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell 158(2):412–421
9. Cummings M, Breitling R, Takano E (2014) Steps towards the synthetic biology of polyketide biosynthesis. FEMS Microbiol Lett 351(2):116–125
10. Fitzgerald JT, Ridley CP, Khosla C (2011) Engineered biosynthesis of the antiparasitic agent frenolicin B and rationally designed analogs in a heterologous host. J Antibiot 64(12):759
11. Medema MH, Breitling R, Bovenberg R, Takano E (2011) Exploiting plug-and-play synthetic biology for drug discovery and production in microorganisms. Nat Rev Microbiol 9(2):131
12. Medema MH, Takano E, Breitling R (2013) Detecting sequence homology at the gene cluster level with MultiGeneBlast. Mol Biol Evol 30(5):1218–1223
13. Newman DJ, Cragg GM (2007) Natural products as sources of new drugs over the last 25 years. J Nat Prod 70(3):461–477
14. Seyedsayamdost MR, Clardy J (2014) Natural products and synthetic biology. ACS Synth Biol 3(10):745–747

15. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147(1):195–197
16. Thanapipatsiri A, Claesen J, Gomez-Escribano JP, Bibb M, Thamchaipenet A (2015) A Streptomyces coelicolor host for the heterologous expression of Type III polyketide synthase genes. Microb Cell Fact 14(1):145
17. Weber T, Rausch C, Lopez P, Hoof I, Gaykova V, Huson D, Wohlleben W (2009) CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. J Biotechnol 140(1):13–17
18. Wolf T, Shelest V, Nath N, Shelest E (2015) CASSIS and SMIPS: promoter-based prediction of secondary metabolite gene clusters in eukaryotic genomes. Bioinformatics 32(8):1138–1143
19. Ziemert N, Ishida K, Liaimer A, Hertweck C, Dittmann E (2008) Ribosomal synthesis of tricyclic depsipeptides in bloom-forming cyanobacteria. Angew Chem 47(40):7756–7759