https://doi.org/10.33451/florafauna.v30i1pp163-166

FLORA AND FAUNA
2024 Vol. 30 No.1 PP 163-166

ISSN 2456 - 9364 (Online)
ISSN 0971 - 6920 (Print)

Identification of biosynthetic gene clusters using bioinformatic tools

Shantirani Thokchom[1], *Saikat Mukherjee[1], Debananda S Ningthoujam[1] and Sumita Banerjee[2]

[1]Department of Biochemistry,
Manipur University, MANIPUR - 795003, INDIA
[2]Department of Oral Pathology,
Dental College, Regional Institute of Medical Sciences,
IMPHAL- 795004 (MANIPUR), INDIA.
*Corresponding Author
E-mail : mukherjeesaikat333@gmail.com

Received : 25.02.2024; ***Accepted*** : 14.03.2024

ABSTRACT

Biosynthetic gene clusters (BGCs) are genomic regions responsible for producing natural products with diverse biological activities. Identifying and characterizing these gene clusters is crucial for understanding the biosynthesis of secondary metabolites and for drug discovery efforts. In recent years, bioinformatics tools have played a pivotal role in the identification, annotation, and analysis of BGCs in microbial genomes. Tools such as antiSMASH, PRISM, and MultiGeneBlast leverage computational algorithms, comparative genomics, and machine learning techniques have been developed to predict BGCs based on the presence of biosynthetic enzymes and other conserved features. These tools enable the inference of chemical structures of natural products encoded by BGCs, further enhancing our understanding of secondary metabolite biosynthesis. They have become indispensable in the field of natural product discovery, empowering researchers to uncover novel secondary metabolites with potential therapeutic applications.

Figure : 00 References : 16 Table : 00

KEY WORDS : AntiSMASH, Biosynthetic gene clusters (BGCs), Multigeneblast PRISM, Secondary metabolites,

Introduction

Biosynthetic gene clusters (BGCs) are groups of genes responsible for producing various natural products, including antibiotics, antifungal, and anticancer agents. Identifying these gene clusters is crucial for understanding the biosynthesis of these compounds and for drug discovery efforts. Biosynthetic tools play a significant role in the identification of biosynthetic gene clusters.

Some of the key roles played by Biosynthetic tools are:

1. **Genome Mining:** Biosynthetic tools aid in the mining of genomic data to identify putative biosynthetic gene clusters. This involves the computational analysis of DNA sequences to detect regions that encode enzymes involved in biosynthetic pathways. Tools such as antiSMASH (antibiotics & Secondary Metabolite Analysis Shell) are commonly used for this purpose. AntiSMASH predict BGCs in microbial genomes and provide detailed annotations of the biosynthetic along with their predicted products[8].
2. **Comparative Genomics:** Comparative genomics tools are employed to analyze similarities and differences in gene content and organization across multiple genomes. By comparing the genomes of closely related organisms, researchers can identify conserved regions that are likely to be involved in the biosynthesis of specific compounds. Tools like MultiGeneBlast facilitate the comparison of multiple genomes to identify conserved gene clusters associated with biosynthetic pathways[10].
3. **Machine Learning and Artificial Intelligence:** Machine Learning and AI techniques are increasingly being applied to the identification of biosynthetic gene clusters. These methods can learn patterns from large datasets of known gene clusters and then use this knowledge to predict the presence of similar clusters in new genomic data. For example, DeepBGC utilizes deep learning algorithms to predict BGCs directly from DNA sequence data, achieving high accuracy in identifying novel biosynthetic gene clusters [15].
4. **Functional Analysis:** Functional Analysis tools are used to characterize the predicted biosynthetic gene clusters and their products. This may involve predicting the chemical structure of the compounds

encoded by the gene clusters, as well as identifying the enzymatic activities involved in their biosynthesis. Tools such as PRISM (Prediction Informatics for Secondary Metabolomes) integrate various bioinformatics approaches to predict the chemical structures of natural products based on their biosynthetic gene clusters[14].

5. **Experimental Validation:** While computational tools can predict the presence of biosynthetic gene clusters, experimental validation is essential to confirm their existence and characterize their functions. Molecular biology techniques such as gene knockout experiments, heterologous expression, and metabolomics are used to validate the predicted gene clusters and study the biosynthesis of the corresponding natural products.

Types of Biosynthetic tools

1. **SMURF (Secondary Metabolite Unknown Regions Finder):** This is not an instrument but rather a bioinformatics tool designed for the identification of biosynthetic gene clusters (BGCs) within fungal genomes, particularly those responsible for the production of secondary metabolites. It's used for genome mining and prediction of novel biosynthetic pathways in fungi. SMURF relies on the analysis of conserved domains and other features associated with biosynthetic enzymes to identify putative BGCs[5].

Working of SMURF

a) **Genome Annotation:** SMURF starts with the annotation of fungal genomes. This involves predicting genes and their functions within the genome sequences.

b) **Identification of Biosynthetic Gene Clusters:** SMURF then scans the annotated genomes to identify regions that contain genes encoding enzymes typically associated with secondary metabolism. These enzymes may include polyketide synthases (PKSs), non-ribosomal peptide synthetases (NRPSs), terpene synthases, and other biosynthetic enzymes.

c) **Cluster Prediction:** Based on the presence of these biosynthetic enzymes and associated domains, SMURF predicts the boundaries of putative biosynthetic gene clusters. It identifies clusters of colocalized biosynthetic genes that are likely involved in the production of secondary metabolites.

d) **Annotation and Visualization:** SMURF provides annotations for the predicted gene clusters, including the types of biosynthetic enzymes present and their putative products. It also offers visualization tools to explore the genomic context of the predicted clusters.

e) **Functional Analysis:** Researchers can further analyze the predicted gene clusters to infer the chemical structures of the corresponding secondary metabolites and investigate their potential biological activities.

2. **AntiSMASH (Antibiotics & Secondary Metabolite Analysis Shell):** It is a widely used bioinformatics tool for the identification and analysis of biosynthetic gene clusters (BGCs) responsible for the production of secondary metabolites in microbial genomes. It provides comprehensive annotations and predictions regarding the types of secondary metabolites encoded by these clusters, as well as the biosynthetic pathways involved. It has been widely used in the field of natural product discovery and synthetic biology. It has contributed to the identification of numerous novel secondary metabolites and the elucidation of their biosynthetic pathways across a wide range of microbial species [2,10,16].

3. **NP Searcher:** It is a bioinformatics tool developed for the identification and analysis of natural product biosynthetic gene clusters (BGCs) within microbial genomes. It utilizes a combination of sequence similarity search, hidden Markov models (HMMs), and machine learning techniques to predict BGCs involved in the biosynthesis of secondary metabolites[7].

4. **CASSIS-SMIPS (Cluster Assignment by Islands of Sites using Single Molecule Real-Time Sequencing)**: This bioinformatics tool was developed for the identification and analysis of biosynthetic gene clusters (BGCs) using Single Molecule Real-Time Sequencing (SMRT) sequencing data. SMRT sequencing, provided by Pacific Biosciences (PacBio), offers long read lengths that can span entire BGCs, making it particularly useful for the comprehensive characterization of these genomic regions.

Gene cluster detection tools that go beyond metabolic gene clusters : While many gene cluster detection tools primarily focus on identifying metabolic gene clusters, some also go beyond this scope to detect other types of gene clusters, such as those involved in the regulation of gene expression, chromatin organization, or other biological processes. Few examples of gene cluster detection tools that encompass a broader range of genomic features:

1. **MultiGeneBlast**: It is a tool for comparative genomics that allows users to search for homologous gene clusters across multiple genomes[11].

2. **CORASON (ClusteR Analysis of Secondary

Metabolite On Nonribosomal peptides): It is a tool for the exploration and analysis of biosynthetic gene clusters in microbial genomes, including both non-ribosomal peptides and other secondary metabolites[3].

3. **CASSIS (Computer Assisted Strain-specific Identification of Secondary Metabolism):** This tool is used for the automated identification and annotation of biosynthetic gene clusters in bacterial genomes, including both known and novel clusters[12].

4. **CMG-Biotools (Clustering of Metagenomic Biotopes and Metabolites for Gene Banks Exploration):** It is a suite of bioinformatics tools designed for the exploration of metagenomic data, including the identification of biosynthetic gene clusters associated with diverse ecological niches[4].

Supplemental resources for metabolic gene cluster prediction methods for predicting substrate specificity: Predicting substrate specificity of biosynthetic enzymes within metabolic gene clusters is a challenging yet crucial task in natural product discovery. While many gene cluster prediction methods focus on identifying the presence of biosynthetic gene clusters, fewer specifically address substrate specificity prediction. However, there are some resources and methods available that can provide supplemental information or aid in the prediction of substrate specificity. Few examples are:

a) **PRISM** (Prediction Informatics for Secondary Metabolomes): It integrates various bioinformatics approaches to predict the chemical structures of natural products based on their biosynthetic gene clusters. It can provide insights into the substrate specificity of biosynthetic enzymes by predicting the chemical structures of the corresponding natural products[14].

b) **NRPSpredictor2:** This is a tool specifically designed for the prediction of substrate specificity in non-ribosomal peptide synthetases (NRPSs). It uses machine learning algorithms trained on known NRPS substrate specificities to predict the substrates of NRPS adenylation (A) domains[13].

c) **SBSPKS (Substrate-Binding Specificity Prediction for Ketosynthase Domains):** SBSPKS is a tool for predicting the substrate specificity of Ketosynthase (KS) domains in type I polyketide synthases (PKs). It utilizes a machine learning approach to predict the substrate specificity based on the sequence and structural features of KS domains[6].

d) **NRPS-PKS**: It is a database that provides information on substrate specificity profiles for adenylation (A) domains in NRPSs and ketosynthase (KS) domains in PKSs. It can serve as a supplemental resource for predicting substrate specificity based on known patterns in these domains[1].

Conclusion

Bioinformatics tools such as AntiSMASH, PRISM and MultiGeneBlast have revolutionized the field of natural product discovery by enabling the rapid and accurate identification of biosynthetic gene clusters in microbial genomes. These tools utilize a variety of computational methods, including sequence similarity search, hidden Markov models, and machine learning, to predict BGCs based on the presence of biosynthetic enzymes. Furthermore, these tools provide comprehensive annotations for the predicted gene clusters, including the types of biosynthetic enzymes present, domain architectures, and putative products. This information allows researchers to explore the biosynthetic potential of microbial genomes and prioritize BGCs for experimental validation and downstream analysis.

In addition to predicting BGCs, bioinformatics tools also play a crucial role in inferring the chemical structures of natural products encoded by these clusters. By integrating genomic data with chemical knowledge databases, tools such as PRISM can predict the chemical structures of secondary metabolites based on the biosynthetic enzymes present in the gene clusters. Overall, bioinformatics tools have become indispensable in the field of natural product discovery, providing researchers with powerful resources for exploring the biosynthetic diversity of microbial genomes and uncovering novel secondary metabolites with potential therapeutic applications.

References

1. Anand S, Prasad MVR, Yadav G, Kumar N, Shehara J, Ansari MZ, Mohanty D. “SBSPKS: structure based sequence analysis of polyketide synthases.” *Nucleic Acids Research*. 2010; **38**(2): 487-496.

2. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, Weber T. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Research*. 2019; **47**(1): 81-87.

3. Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland BLC, Mavrommatis K. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell*. 2014; **158**(2): 412-421.

4. Goncalves C, Briand M, Araujo D, Nunes BS, Previati M, Silva LJ, Andreote FD. CMG-biotools, a free workbench for basic comparative microbial genomics. *PloS one*. 2015; **10**(11): 0134803.

5. Khaldi N, Seifuddin FT, Turner G, Haft D, Nierman WC, Wolfe KH, Fedorova ND.SMURF: genomic mapping of fungal secondary metabolite clusters. *Fungal Genetics and Biology*. 2010; **47**(9): 736-741.

6. Khayatt BI, Overmars L, Siezen RJ, Francke C. Classification of the adenylation and acyl-transferase activity of NRPS and PKS systems using ensembles of substrate specific hidden Markov models. *PloS one*. 2013; **8**(10): 62136.

7. Li MH, Ung PMU, Zajkowski J, Garneau-Tsodikova S, Sherman DH. Automated genome mining for natural products. *BMC Bioinformatics*. 2009; **10**(1): 185.

8. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research*. 2011; **39**(2): 339-346.

9. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research*. 2011; **39**(2): 339-346.

10. Medema MH, Takano E, Breitling R. Detecting sequence homology at the gene cluster level with MultiGeneBlast. *Molecular biology and evolution*. 2013; **30**(5): 1218-1223.

11. Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, Moore BS. Minimum information about a biosynthetic gene cluster. *Nature Chemical Biology.* 2015; **11**(9): 625-631.

12. Navarro- Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, Medema MH. A computational framework to explore large-scale biosynthetic diversity. *Nature Chemical Biology.* 2020; **16**(1): 60-68.

13. Röttig M, Medema MH, Blin K, Weber T, Rausch C, Kohlbacher O. NRPSpredictor2-a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Research*. 2011; **39**(2): 362-367.

14. Skinnider MA, Dejong CA, Rees PN, Johnston CW, Li H, Webster AL, Magarvey NA.Genomes to natural products Prediction Informatics for Secondary Metabolomes (PRISM). *Nucleic Acids Research*. 2015; **43**(20): 9645-9662.

15. Skinnider MA, Merwin NJ, Johnston CW, Magarvey NA, Global Natural Product Social (GNPS) Networking. PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Research*. 2019; **47**(1): 273-276.

16. Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, Lee SY. AntiSMASH 3.0- a comprehensive resource for the genome mining of biosynthetic gene cluster. *Nucleic Acids Research*. 2015; **43**(1): 237-243.