

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900

Open access books available

185,000

International authors and editors

200M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Multi-Omics Data Mining: A Novel Tool for BioBrick Design

Angie Burgos-Toro, Martin Dippe, Andres Felipe Vásquez, Eric Pierschel, Ludger Aloisius Wessjohann and Miguel Fernández-Niño

Abstract

Currently, billions of nucleotide and amino acid sequences accumulate in free-access databases as a result of the *omics* revolution, the improvement in sequencing technologies, and the systematic storage of shotgun sequencing data from a large and diverse number of organisms. In this chapter, multi-*omics* data mining approaches will be discussed as a novel tool for the identification and characterization of novel DNA sequences encoding elementary parts of complex biological systems (BioBricks) using *omics* libraries. Multi-*omics* data mining opens up the possibility to identify novel unknown sequences from free-access databases. It also provides an excellent platform for the identification and design of novel BioBricks by using previously well-characterized biological bricks as scaffolds for homology searching and BioBrick design. In this chapter, the most recent mining approaches will be discussed, and several examples will be presented to highlight its relevance as a novel tool for synthetic biology.

Keywords: genome, transcriptome, proteome, data mining, metabolic pathway, BioBricks design, multi-*omics*, synthetic biology

1. Introduction

1.1 The omics revolution

Within the last decades, a magnificent transformation in biology took place when a huge success in sequencing, bioinformatics, and bioanalytics was achieved. Several technologies were created to decrypt the metabolism of cells or interactions within tissues, organisms, and even entire ecosystems based on the identification of genes (genomics), mRNA (transcriptomics), proteins (proteomics), and metabolites (metabolomics) [1]. Since the discovery of the DNA structure by Watson and Crick in 1953 [2], an ever-increasing number of technologies for gene identification and characterization was established. One of the most relevant breakthroughs in DNA characterization was the invention of Sanger's sequencing in 1977 [3]. This sequencing technique uses chemical analogs of the deoxyribonucleotides (dNTPs, monomers of DNA strands) called dideoxynucleotides (ddNTPs), which lack the 3' hydroxyl group that is required for extension of DNA chains and therefore cannot form a bond with the 5' phosphate of the next dNTP [4]. The overall advantages of

accuracy, robustness, and ease of use against other established methods led Sanger sequencing to become one of the most common technologies used to sequence DNA. Several improvements were subsequently applied to this technique, such as the use of fluorometric detection and capillary-based electrophoresis, thus contributing to the development of automated DNA sequencing machines [5–11]. These machines allowed researchers to obtain sequence reads slightly less than one kilobase (kb) in length and boosted the development of other crucial technologies such as the Polymerase Chain Reaction (PCR) in 1985 and the recombinant DNA technology in the following years [12, 13].

In parallel to the development of large-scale dideoxy sequencing methods, a new technique set the novum for next-generation DNA sequencers. This approach remarkably varies from the abovementioned methods as it does not involve the use of radio- or fluorescently labeled dNTPs. Instead, it is based on a luminescent method for measuring pyrophosphate synthesis in a process called pyrosequencing [14]. This sequencing technology is a two-enzyme process starting with the conversion of pyrophosphate into ATP (by an ATP sulfurylase) and the subsequent use of ATP as a substrate for luciferase, thus emitting light proportional to the amount of pyrophosphate available. Pyrosequencing became a popular technique for two major reasons: (i) it uses natural nucleotides instead of modified ones, and (ii) that sequencing results can be obtained in real-time without requiring time-consuming electrophoresis. In addition to pyrosequencing, other sequencing technologies were also devolved - the most important probably being the Solexa method, later acquired by the company Illumina [15]. Hereby, adapter-bracketed DNA molecules pass a lawn of complementary oligonucleotides bound to a flow cell. This method involves solid-phase PCR with neighboring clusters of clonal DNA strands in a process called “bridge amplification” [15–17]. Apart from Illumina, which is probably the most important technique currently in use, other sequencing companies established their novel methodologies [18, 19], which are known as the second-generation sequencing techniques. The most notable second-generation sequencing platform is probably Ion Torrent. It is the first “post-light sequencing” technology with neither using fluorescence nor luminescence. Its methodology is based on beads bearing clonal populations of DNA fragments washed over a pico well plate, thereby releasing protons measured via the generated pH difference [20].

Recently, a third sequencing generation started with the invention of S. Quake in 2003 termed Single Molecule Sequencing (SMS) [21, 22]. Its principle is similar to Illumina but skipping bridge amplification. In SMS, DNA templates attached to a planar surface and propriety fluorescent reversible terminator dNTPs (dubbed as “virtual terminators”) are washed over one base at a time and imaged, before cleavage and cycling the adjacent base over. SMS has been recently improved in the Single-Molecule Real-Time (SMRT) platform from Pacific Biosciences, available for the PacBio machines [23]. During SMRT runs, DNA polymerization happens in arrays of microfabricated nanostructures called zero-mode waveguides (ZMWs) which are essentially tiny holes in a metallic film covering a chip. It allows visualization of single fluorophore molecules because the zone of laser excitation is so small that it allows distinction over the background of neighboring molecules in the solution [24]. Nonetheless, the probably most anticipated third-generation DNA sequencing method is nanopore sequencing which enables researchers to detect and quantify all types of biological molecules [25]. Its principle was theoretically established even before second-generation sequencing emerged by demonstrating that single-stranded RNA or DNA could be driven across a lipid bilayer through a large α -hemolysin ion channel by electrophoresis. Furthermore, passage through the channel blocks ion flow, decreasing the current for a length of time proportional to the length of the nucleic acid [26]. With Oxford Nanopore Technologies (ONT)

as the first provider of nanopore sequencers and their nanopore platforms GridION and MinION [27, 28], the latter of which is a small, mobile phone-sized USB device (released in 2014) [29]. Despite the admittedly poor quality profiles currently observed, it is hoped that such sequencers represent a genuinely disruptive technology in the DNA sequencing field in the future, producing incredibly long read (non-amplified) sequence data far cheaper and faster than what was previously possible [28, 30]. The average read length, error rate, total number of reads, and run prices vary significantly among the different sequencing methodologies. Thus, the selection of the appropriate technology for sequencing is a crucial step that depends on the purpose of the study. For instance, Illumina and Ion Torrent produce accurate short reads ideal for the analysis of fragmented DNA, while PacBio and MinION produce long reads with a lower accuracy but very useful, for example, for the assembly of scaffolds during genome sequencing.

Similar to the development of advanced techniques for sequencing nucleic acids, other methods have been extensively developed for dissecting the proteome [31] and metabolome [32] of a multitude of organisms. Of these omics approaches metabolomics, however, is distinct from the others. In metabolomics not a set of linear (1D) molecules with a sequence of defined monomers (4 bases or 21 amino acids) is to be determined, but a wild bunch of different 3D compounds. Eventually, a large number of databases have been developed to collect all these information, which provide excellent platforms for data mining as will be discussed in the following chapters.

2. Genome and transcriptome data mining

The exponential accumulation of data in genomic databases during the last decades has motivated the creation of bioinformatics tools to explore, relate and understand the genetic information from a vast number of organisms [33, 34]. These bioinformatics tools have been validated by experimental data, thus strengthening the design and assembly of novel biological entities (i.e., genes, RNA molecules, proteins, and metabolites). Those biological entities that can be used as building blocks for the assembly of artificial biosynthetic pathways are known as BioBricks. Consequently, the selection and design of BioBricks is important to further create and understand complex biological systems and biofactories of relevance in industrial biotechnology [35]. The general idea of comparing genomic sequences to identify such novel components of different metabolic pathways is not new. In fact, early in the 1970s, several efforts were performed to elucidate physiological and metabolic information through the comparative analysis of genetic sequences [36–38]. Classical genetics and reverse genetics approaches were then used to identify, annotate, compare, and connect genetic clusters associated with biosynthesis, using previously reported genetic data sets [39, 40].

It was not until 1999 that Genome Mining (GM) formally emerged as a strategy for the computational analysis of genetic sequences that sought to recognize patterns between them within the framework of the human genome project. Later, alongside bioinformatics advances in the area of microbiology, GM acquires new attributes, building the concept known today: a bioinformatics approach that aims to predict DNA sequences associated with physiological and/or metabolic events, allowing the elucidation/prediction of metabolic pathways that lead to secondary metabolites of scientific and industrial interest [35, 38, 41, 42]. Today, GM is not limited only to genomic predictions but seeks a holistic approach that includes the entire spectrum of molecular biology, articulating the prediction of the products of gene expression, the control of that expression, as well as the identity and structure of those potential

metabolites, strengthening the creation of biological models that allow the comparison, understanding, and manipulation of cellular molecular systems [41, 43].

GM was initially developed in bacterial models and demonstrated a high relevance for synthetic biologists and metabolic engineers, thus becoming one of the biggest breakthroughs in molecular biology and biotechnology [38, 44]. Between the 1990s and 2000s, the genus *Streptomyces* (which is well known for its production of valuable antibiotics) was extensively studied at the experimental level, which allowed the identification of a large number of gene sequences involved in secondary metabolite production, regulation and antibiotic resistance. Comparison of gene sequences between different species of this genus, revealed a total of about 30 Biosynthetic Gene Clusters (BGCs) associated with the biosynthesis of such secondary metabolites [45, 46]. Following these advances, GM was extended to study novel bacterial genera with abundant genomic information and was initially used to fight against bacterial resistance [47, 48]. During the last years, GM was successfully used as a tool for the identification of alternative pathways for the biosynthesis of different natural products in diverse microorganisms [33, 49], an approach which usually proved to be more efficient than other screening methods used for the identification of novel enzymes of relevance for the biosynthesis of secondary metabolites [33, 49].

Recently, GM was also scaled up to eukaryotic models, thus revealing that multiple BGCs contain not only relevant information regarding the biosynthesis of secondary metabolites but also valuable information to study evolutionary events and ecological adaptation of different gene clusters [38, 50, 51]. A good example of the vast collection of BGCs predicted up to now can be found on the “Atlas of Biosynthetic Gene Clusters”, a database of the Joint Genome Institute founded in 2015. This Atlas contains data on predicted and experimental gene clusters related to many secondary metabolites. As of June 2021, there are a total of 411,006 biosynthetic gene clusters reported, of which only 1285 have been experimentally validated [52]. GM is completely dependent on bioinformatics and computational technology available for the analysis of a large dataset. Thus, to boost the potential of this information, the development of novel computational tools and algorithms as well as the interest of researchers to join this effort is still required [42, 51]. There are currently a variety of methods for performing GM using the available genomic information that will be further discussed hereafter.

2.1 Classical genome mining

The “classical” form of GM consists of the search for enzymes linked to the synthesis of secondary metabolites, by mining highly conserved sequences [35]. Before the current databases (composed of hundreds of genomic datasets and several bioinformatics tools) were established, novel sequences were evaluated by using reverse genetics, where genomic libraries were scanned for basic biosynthetic genes associated with a metabolic pathway of interest [38, 53]. Those annotations had to be performed manually and by obtaining experimentally corroborated results. This formed the basis of classical GM, which provided the first consensus sequences to be compared with the vast amount of novel sequences obtained from different next-generation sequencing platforms [54]. Both, reverse genetics and GM follow the same mining pattern: one or several reference sequences, whose enzymatic products were already experimentally validated, are used to compare them with the genomes of interest and to identify homologous sequences in the organism of interest. Sequences of interest are considered as being generally associated with catalytic domains and highly conserved motifs [35, 38].

Classical GM was initially focused on the identification of genomic clusters associated with enzymes for the production of secondary metabolites, that involve

the following bacterial groups of enzymes and bioactive peptides: (i) polyketide synthases (PKSs); (ii) non-ribosomal peptide synthetases (NRPSs); and ribosomally and post-translationally modified peptides (RiPPs) [55–57]. Sequence comparison of these groups of proteins allowed the subsequent identification of conserved motifs that are currently helping to identify novel BGCs in pre-existing genomes, without resorting to the strenuous processes of experimentation and first considering the bioinformatic *in silico* approach [58]. Thus, numerous examples have demonstrated the advantage of GM as a successful screening tool for evaluating the ability of one organism to produce a particular metabolite based on the available BGCs information [59–61]. An example of this is presented by Su et al. who performed GM on a strain of *Bacillus subtilis* (i.e., NCD-2), initially predicting its potential for the production of fengicin, surfactin, bacillaene, subtilosin, bacillibactin, bacillosin and other not previously reported molecules, that were later detected by UHPLC-QTOF-MS/MS in its fermentation extracts [62]. The increasing popularity of classical GM promoted the development of GM-specialized databases and novel bioinformatics tools with improved homology searching tools, specialized sequence analyses, and advanced prediction algorithms. A list of some currently available GM specialized databases and related bioinformatics tools are presented in **Tables 1** and **2**, respectively.

Currently, the most popular platform for GM of bacterial and fungal genomes is antiSMASH. It is up to now the most comprehensive by integrating its own database and incorporating different prediction tools [63]. The key of its popularity results from the integration of different complex secondary metabolite-specific gene analysis methods using a much more researcher-friendly interface [82]. Unfortunately, as shown in the tables, most advances have been made in bacteria and there is still a need to improve or create new bioinformatics tools to enable GM in other organisms such as fungi and especially plants, which commonly do not have biosynthetic gene clusters but a separated, often compartmentalized (cell type specific) synthesis of secondary metabolites, including transport of intermediates between cell types and even organs [83, 84].

2.2 Comparative genome mining

Classical GM alone fails to identify BGCs in genomic regions that do not follow a classical modular gene topology, as described by Donadio et al. since 1991. The

Database	Description	Ref.
antiSMASH database	Comprehensive resource on BGCs for secondary metabolites identified in bacterial genomes.	[63]
BACTIBASE	Open-access database used for the characterization of bacterial antimicrobial peptides.	[64]
ClusterMine360	Contains over 200 curated entries of BGCs clusters including classification of the potential compounds produced, taxonomic information of the producing organisms, and links to original data.	[65]
CSDB/r-CSDB	Manually curated database containing more than 160 PKS, NRPS, and PKS/ NRPS BGCs.	[66]
DoBISCUIT	Contains a literature-based collection of BGCs for PKS and NRPS.	[67]
IMG-ABC	Contains automatically identified gene clusters, clusters with known biosynthesis products, and secondary metabolites.	[68]

Table 1.
Main databases focused on biosynthetic gene clusters (BGCs) encoding secondary metabolites.

Tool	Description	Ref.
antiSMASH	Fully automated tool for extracting genome data from bacteria and fungi to search for BGCs.	[69]
BiG-SCAPE	Uses the distance between BGCs (identified with antiSMASH), to create sequence similarity networks.	[70]
CLUSEAN	Allows homology searches and identification of conserved domains in BGCs of genes encoding for PKS and NRPS. Also classifies enzymes and predicts the domains specificity.	[71]
CLUSTER FINDER	Uses a probability approach to recognize BGCs in genomic and metagenomic data.	[72]
EvoMining	Uses phylogenetics to recognize, compare and identify BGCs associated with primary metabolism but that present a divergent phylogeny.	[73]
FunGeneClusterS	Allows the prediction of BGCs based on genomic and transcriptomic data for fungi.	[74]
MIPS-CG	Allows the identification of totally new BGCs using only genomic data.	[75]
NaPDoS	Detects and analyze genes associated with secondary metabolites.	[76]
PhytoClust	Detects BGCs of secondary metabolites in plant genomes.	[77]
PKMiner	Predicts novel BGCs of type II PKS and aromatic polyketide chemotypes using their conserved aromatase and cyclase domains.	[78]
plantiSMASH	An antiSMASH' version that uses plant genomes.	[79]
SBSPKS	Allows chemical analysis of experimentally characterized BGCs for PKS/ NRPS proteins.	[80]
SMURF	Used for mining BGCs in fungi to identify conserved domains in PKS, NRPS, PKS/NRPS hybrids, and terpenoid genes.	[81]

Table 2.
Main tools for mining secondary metabolite biosynthesis gene clusters.

organization of open reading frames (ORFs) associated with secondary metabolite-producing genes that generally follow an order of distribution between catalytic and structural domains for modular PKSs or NRPSs, for example, is called a modular pattern [39]. These extensively described and annotated modules serve as a template for comparison with new sequences from available genomes [42].

Leblond and coworkers found more than 3300 BGCs for about 16,500 possible NRPS-associated enzymes in *Streptomyces ambofaciens*. However, when evaluating the potential enzymes *in silico*, they realized that many did not follow the modular pattern used as a template [85]. This, indeed, reduced the possibilities of modeling the possible secondary metabolites that could be produced by this bacterium. This is certainly an example of the current limitations of classical GM, which must contemplate new technologies (e.g., artificial intelligence (AI) and machine learning (ML)) in response to unconventional sequences that do not completely follow the expected organization.

One way to address these limitations is by integrating already existing tools that are focused more on the identification of patterns related to phylogeny and evolution instead of molecular function. For example, descriptions of lineage relationships can be made and some non-modular combinations of putative BGCs can be described between organisms that may not belong to the same taxonomic level. These results are not only valuable for the search for pathways to new natural products, but they also allow evolutionary reconstruction in the creation of metabolic pathways that respond to defense, competition, and attack of organisms in their ecosystem [86]. In plant metabolomics, such phylogentic relationships based

on an untargeted fingerprint approach of natural products of different species were for the first time described in 2013 for *Urtica* species [87], still awaiting a full correlation with genomic data.

Two different ways of using phylogenetics approaches for comparative GM can be defined: In the first one, phylogenetics trees are constructed using both the whole sequences of the organisms under study and a pool of conserved well-characterized gene clusters associated to the production of a defined compound. In this way, BGC lineages can be traced and evolutionary relationships between apparently unrelated organisms can be established. Abdelmohsen et al. used this strategy to investigate biosynthetic pathways in actinomycetes isolated from marine sponges from the Red Sea. After a combination of taxonomic evaluation using the 16S ribosomal gene, PCR amplification of genes associated with modular PKS and NRPS, and phylogenetic analysis, the authors found that 20 of the actinomycetes isolates (speeded over 10 genera) possessed at least one of the biosynthetic genes analyzed [88]. This method has been extensively applied to identify novel potential BGCs [70, 89] and to create new gene clusters that can be further related to already annotated genomes of organisms previously studied at the experimental level.

The use of comparative GM has also allowed the identification of genes involved in the production of secondary metabolites in bacteria, by considering horizontal gene transfer events and phylogenetic analysis. Here, relationship trees are constructed using genes that are directly associated with the creation of specific compounds/secondary metabolites [90]. In this model, gene relationships are inferred primarily using the biosynthetic gene sequences only, and later those relationships are contrasted or strengthened by evaluating the rest of the organism's genome [91]. An example of the use of this method are studies conducted on the genus *Streptomyces*, where the production of secondary metabolites was again evaluated considering events of lateral gene transfer. It was found that, although horizontal gene transfer of the studied BGCs is not so frequent, the transfer of exogenous regulatory, resistance, and secondary metabolite production genes can significantly contribute to recombination events in those BGCs. Thus, comparative GM brings new relevant concepts such as the variable nature of those BGCs and their diversification even within very specific levels of phylogenetic discrimination. This undoubtedly paves the way not only to understand the evolution of BGCs in microorganisms but also to understanding the ecological landscape that it influences [91].

Currently, one of the methods to specifically evaluate putative catalytic domains in enzymes, using phylogenetic algorithms, is the Natural Product Domain Seeker (NaPDos), which organizes sequences into clades and allows the recognition of lineages of organisms capable of producing selected metabolites [76, 92]. This represents a new approach for the evaluation of possible non-homologous and undescribed enzymes (shown for modular PKS and NPRS) and to elucidate new chemical structures not yet identified. NaPDos initially contained only data from PCR fragments but now is a comprehensive tool that also includes genomics and metagenomics data [93]. This is particularly important because it allows the evaluation of genomic data obtained from complex samples such as soils, sediments, water sources, wastes, etc. (metagenomics). With NaPDos it is even possible to estimate the diversity of microorganisms from the sampled source, as well as to evaluate the genetic potential for the biosynthesis of different metabolites [93].

2.3 Genome mining in synthetic biology

The identification of novel BGCs resulting from genomic mining studies represents a great opportunity for synthetic biologists and metabolic engineering as it

allows the identification, construction, synthesis, and expression of BioBricks in heterologous models or to discover natural compounds with outstanding properties. One of the most significant commercial examples of this application has been observed during the engineering of yeast for the biosynthesis of valuable products such as artemisinin (an antimalarial drug) by using BioBricks identified through GM [35, 94]. Recently, GM has been also used to identify more than 70 syntheses involved in the production of hypermodified peptide cytotoxins (i.e., unique, and valuable chemotherapeutics) by mining prokaryotic diversity [95]. With the help of GM, the identification of several cryptic metabolic pathways has been possible, giving way to combinatorial biosynthesis, which can be used in the construction of biosynthetic units, following the pattern of BGCs. These approaches also present challenges mainly related to our current understanding of the interdependent metabolic circuits, and the complexity in tracking them. This will certainly require many more efforts from bioinformatics to enrich genomic mining by including additional omics data such as transcriptomics, metabolomics, and proteomics not only for microorganisms but also for eukaryotes with their complex, usually unclustered biosynthetic production networks [96].

2.4 Transcriptome mining

A transcriptome represents a “snapshot” of a RNA population in a certain tissue or at a specific developmental stage. Compared to the genomic information of the same organism, a transcriptomic dataset is less complex as it does not contain any information, for example, on the untranslated regions of a genome (e.g., promoters). Transcriptomes also do not provide information on the physical organization of the individual genetic elements—a fact which in turn represents an obstacle for the application of classical GM methods (see previous sections) used, for instance, for pathway elucidation in plants. However, several advantages make transcriptome mining (TM) a valuable alternative in the last years: First, unlike in a “static” genome, differential analysis is possible for transcriptomic data. Thus, the identification of tissue-specific transcripts (pathways restricted to special organs) and discrimination of non-functional RNAs (pseudogenes) is much easier than in GM approaches. Secondly, the less complex datasets facilitate mining in organisms with large and complex genomes such as plants [97], which in general developed multi-member gene families with redundant functions during evolution. In conjunction with the fact that the organization of biosynthetic pathways into gene clusters is exceptional in plants [98], TM is increasingly used in this class of organisms to mine for NP pathways as well as to study different aspects of plant physiology. Recent examples for the latter purpose include the dissection of the response to changing temperatures [99], drought stress [100], or defense against pathogens in model and non-model plants [101, 102].

First reports on TM used for the discovery of NP biosynthetic genes date back to the first decade of the 21st century. The reports were based on so-called expressed sequence tag (EST) databases [103], which were developed as an alternative to earlier microarray-driven methods for expression analysis. Milestones for the application in the plant field were the establishment of specific EST databases [104] and the access to programs that used both microarray data and transcriptome datasets in the frame of transcriptome profiling (e.g., eVOC [105]). Continued software development led to more advanced approaches which integrated data modeling in targeted plant engineering [106]. Alongside with the use of co-expression analysis as a standard tool in multifaceted mining strategies [107] and the current decrease in prices for transcriptome sequencing, the developments led to a continuous increase in the annual output of TM-based publications (3 in 2003, 84 in 2020).

For instance, all classes of NPs found in plants were targeted using TM in the last years. Most reports focused on **terpenoids**, including papers on the identification of single enzymes such as terpene cyclases/synthases [108], associated biocatalysts [109] or comparative evolutionary studies of genes in whole plant families such as Pinaceae [110] or Lamiaceae [111]. An outstanding example is the mining for biocatalysts involved in the biosynthesis of the insecticidal limonoid azadirachtin in neem (*Azadirachta indica*) [112]. By using a comparative analysis of three limonoid-containing species from the order Sapindales, the authors could identify key enzymes involved in the early steps of the pathway, namely the initial terpene cyclase forming the basal triterpene scaffold and subsequent cytochromes involved in tailoring modifications. In the field of **alkaloids**, TM was similarly applied, yielding the enzyme norbelladine synthase from *Narcissus pseudonarcissus* [113]. This enzyme, which is used for a coupling step during the synthesis of the anticancer agent galantamine in *Narcissus* species, was fished by a TM-based screening for functional homologs of an enzyme catalyzing a similar enzymatic reaction in opium poppy. Hagel and co-workers [114] used a similar but broader approach to compare plants with a pronounced production of benzyloquinoline alkaloids. Differential analysis of the transcriptomes and metabolomes of 20 species from the order Ranunculales revealed 850 genes that are potentially involved in alkaloid biosynthesis and are interesting candidates for use in alkaloid Synthetic Biology. A noteworthy example concerning the biosynthesis of plant **phenolics** is the study of Lau and Sattley [115], which describes mining for enzymes required for the production of podophyllotoxin. This lignan is an antiviral polyphenol isolated from mayapple (*Podophyllum peltatum*), and six of the enzymes involved in its biosynthesis could be identified by TM followed by subsequent co-expression in tobacco. Another example is the insight from TM and Metabolomics in the synthesis of hypericin in the medicinal plant St. John's wort (*Hypericum perforatum*) [116].

Future studies will certainly use extensive TM to further explore the biosynthetic machineries to high-value metabolites other than terpenes, alkaloids, and phenolics. In agreement with this assumption, the latest reports on TM already target pathways to antimicrobial cyclopeptides [117], polysaccharides [118], or compounds derived from fatty acids [119]. In general, TM studies will definitely benefit from the integration of multi-level omics data in the future. Such comprehensive methods have already been applied in proof-of-concept studies, including the combination of TM with proteomics to mine for cyclopeptides [120] or in-plant "regulomics", i.e., in software tools comparing transcriptomes with (epi)genomic data to identify regulatory networks [121].

3. Metabolic data mining

Metabolism is typically defined as the sum of pathways and cycles representing all the sets of biochemical reactions occurring at a cell and in which the product of a particular chemical reaction becomes the substrate of the subsequent reaction [122]. Certainly, the understanding of this concept is key in the realm of biological sciences, especially in the post-genomic era, where we have embraced a paradigm shift from a gene-centered view to an increasing interest in omics-driven high-throughput data types, sources, and approaches [123]. In line with the current move towards systems biology, the mining of metabolism data (metabolic data mining) includes not only the systematic study of component metabolites (i.e., **metabolomics**) [124], but also of all the controlled biochemical reactions in an organism responsible for their production, which is more recently understood under the name of **reactomics** [125] and related processes such as in **fluxomics**

[126, 127]. In metabolomics, numerous subclasses have emerged, as in distinction to especially genomics, a really holistic determination of the metabolome is impossible: no method exists to extract and analyze all metabolites of an organism completely in one experiment. Unlike in genomics, transcriptomics or proteomics, metabolome analytics cannot rely on a one dimensional sequential biopolymer of a limited number of monomer units and a few handful of derivatizations (methylation, post-translational modifications etc.). Instead, most compounds are unique, they are rarely produced by linear monomer assembly processes which can be deconvoluted by standardized processes. But instead a metabolome is a mixture of compounds with highly complex 2D and mostly 3D molecular structures of maximum variability and physicochemical property divergences (e.g., sugars vs. triglycerides). Subclasses have thus emerged, e.g., lipidomics or glycomics. Along with the great advances of computing technologies, all types of studies -especially when applied in combination- have led us to witness an unprecedented revolution in biotechnology by finding patterns or trends that explain the behavior of large data sets in a specific context and as automated as possible. Thus, during the last decade, a large number of metabolic pathways have been mined to identify the key elements and modules for the production of drugs, foods, fuels, and a plethora of bioactive compounds [128–130], including the combination of transcriptome and metabolome studies [116].

The trifold correlation of metabolomic, transcriptomic/genomic and phenotypic data ideally allows to identify both gene loci responsible and the biosynthetic components responsible for a property (phenotype), the biosynthetic pathways for their production, and the genetic control elements associated with them (GWAS—genome wide association study). This allows e.g., improved molecular breeding in plants without the necessity of producing GMOs. An example is a study on downy mildew resistance in hops (*Humulus lupulus*), i.e., tackling its most devastating pathogen by identifying the intrinsic strengths of its chemical defense. The identification of key metabolites responsible for mildew resistance, their associated pathways and genetic breeding markers associated with downy mildew resistance now allows the targeted (non-GMO) molecular breeding of resistant phenotypes [131]. The same tools can, of course, also be used for higher production using genetic improvement (GMOs) [80]. The different strategies for the identification of these metabolic pathways via data collection and coupling, reactome reconstruction, and rational exploration of the chemical space will be further discussed.

3.1 Metabolic data collection and coupling

A typical workflow in metabolic data mining aimed to elucidate interaction networks and reactomes is shown in **Figure 1**. Initially, metabolic data is collected including information on enzymes and metabolites. Then, the recognition and

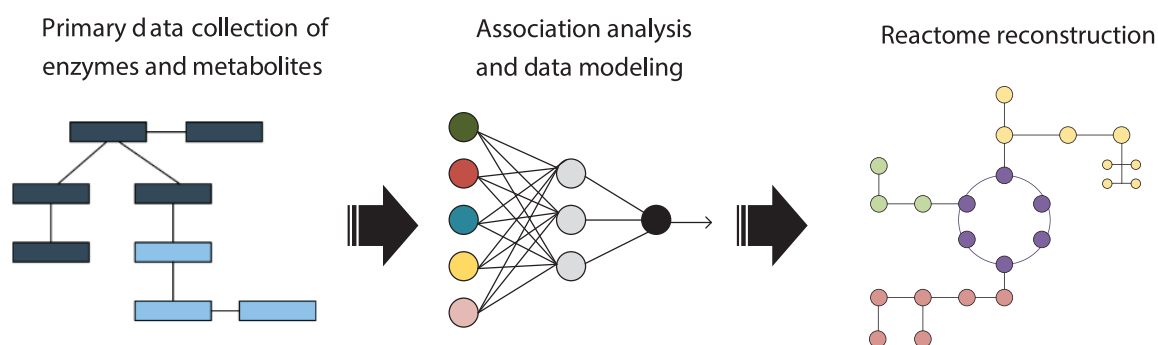


Figure 1.
Standard workflow in metabolic data mining to elucidate interaction networks and reactomes.

coupling of network patterns are carried out by association analysis and data modeling to obtain a reduction in data dimensionality. Finally, reactomes are reconstructed to elucidate the corresponding network dynamics and topology [132]. This knowledge forms the basis for future metabolic engineering experiments aimed to enhance the production of the desired compound or to assemble novel native but also synthetic/unnatural biosynthetic pathways. Interestingly, the current advances in the development of novel BioBricks and the design of novel artificial metabolic networks promote the rapid and efficient coupling of a series of biological parts into a highly reusable large-scale framework [133].

3.2 Proficient exploration of chemical space: natural products and fragments

Metabolic data mining also may involve the use of small compounds derived from the primary and, most especially, secondary metabolism of living organisms. These metabolites, typically referred to as natural products (NPs), have largely been used as a source of chemical entities with promising physicochemical, medicinal or other features, being used directly (unmodified), as a substructure, or as inspiration for a structurally similar chemical scaffold [134, 135]. NPs have been used for ages as medicines than the synthetic bioactives and as scaffolds for the rational design of novel synthetic drugs [136, 137]. Interestingly, they occupy a much larger fraction of the ensemble of all chemical compounds (i.e., have a larger structural diversity), which is classically known among theoretical and computational chemists as **chemical space** ($\sim 10^{60}$ molecules) [138, 139]. In the field of medicinal chemistry, and considering we only know just a bit portion of the estimated chemical space ($\sim 10^8$ molecules) [140], the use of NP-based libraries represents a priceless opportunity for scientists to make bigger and faster leaps within it [141, 142]. This fact represents an additional advantage taking into account that conventional combinatorial chemistry (usually termed combichem) without input from natural products initially had very limited success in novel drug discovery [141, 143], having its strength rather in optimization in most cases [141]. On the other hand, an alternative scenario intended to explore the chemical space more profoundly and, thus, may be used to harness metabolic data involves the principles of molecular fragmentation. According to this technique, a chemical compound of interest is not identified and evaluated as a whole, but instead, it is developed starting from structural molecular components usually within the range 120–300 Da (i.e., fragments) [144, 145]. Although many current chemical libraries are available as fragments per se, various cleavage methods such as RECAP (Retrosynthetic Combinatorial Analysis Procedure) have been widely used to deconstruct chemical libraries of both NPs and other classes of chemical entities [146, 147]. Among the many advantages of using fragments are not only their potential to navigate into the chemical space in a more cost-effective manner compared, for example, to drug-sized molecules, but also their potential to favor the protein-ligand complementarity and facilitate selectivity adjustments during optimization processes (a more detailed description is given in **Figure 2**) [148, 149]. Once more, within the field of BioBricks, the possibility of understanding every fragment as an independent brick could facilitate not only the recovery of specific substructures during a virtual screening (VS) protocol but also the coupling of the best combinations of substructures to obtain a final candidate for further development. It is worth mentioning that fragments could be “recycled” to be considered in the development of a bigger compound if other partner fragments can supply -and balance- particular physicochemical properties of interest. This is fully illustrated in terms of ligand efficiency (LE) metrics as a phenomenon called fragment “rescue” effect [150]. Through an application of these kinds of concepts and approaches, the

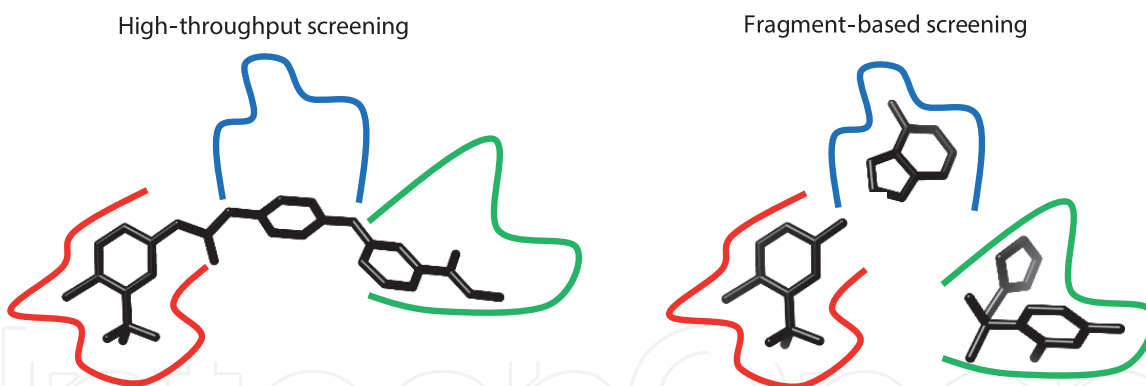


Figure 2. Comparison between typical high-throughput screening and fragment-based screening. In the left panel, it is evident that although one specific part of the drug compound exhibits a good fit within most of the pocket of a hypothetical target protein (red curved line), the other two parts of the same compound do not occupy any specific binding (blue curved line) or occupies subsites of the active center only partially (green curved line). In contrast, the right panel shows that the consideration of fragments for screening allowed the identification of chemical entities with high inherent affinity to the corresponding pockets. Although only shape and size are included in the illustration for clarity, many other physicochemical characteristics such as lipophilicity and charge may affect the complementarity between a chemical moiety and its target receptor.

scientific community may benefit from metabolomic data mining of compounds able to mediate diverse functions in biological systems.

4. Conclusions

Multi-omics data mining has revolutionized science by enabling overlaps among different fields of study such as biochemistry, molecular biology, synthetic biology, organic and medicinal chemistry, computational chemistry, chemical engineering, and high-performance computing. This represents a crucial breakthrough that is expected to accelerate our comprehension of complex biological systems and, most interestingly, the identification, selection, and recovery of novel pieces of biological information in the form of BioBricks for the design of biofactories. Currently, we have unprecedented access to large multi-omics data repositories, which make possible the discovery, identification, and coupling of these BioBricks. This is an important step to unleash different biological functions, or to rationally design metabolic pathways for the biosynthesis of valuable products. However, there is still a need for integrating additional cutting-edge technologies in computing and data science such as machine learning, artificial intelligence, and big and smart data analytics that can further boost the discovery and *de novo* design of BioBricks with high impact in pharma, cosmetics, fine chemical and nutraceutical industries.

Conflict of interest

The authors declare no conflict of interest.

IntechOpen

Author details

Angie Burgos-Toro¹, Martin Dippe², Andres Felipe Vásquez^{3,4}, Eric Pierschel²,
Ludger Aloisius Wessjohann² and Miguel Fernández-Niño^{2*}

1 Department of Biology, Universidad Nacional de Colombia, Colombia


2 Department of Bioorganic Chemistry, Leibniz-Institute of Plant Biochemistry,
Halle, Germany

3 Grupo de Diseño de Productos y procesos (GDPP), Facultad de Ingeniería,
Universidad de los Andes, Bogotá, Colombia

4 Naturalius S.A.S., Bogotá, Colombia

*Address all correspondence to: mfernand@ipb-halle.de

IntechOpen

© 2021 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Hood L, Rowen L. The human genome project: Big science transforms biology and medicine. *Genome Medicine*. 2013;**5**:1-8. DOI: 10.1186/GM483
- [2] Watson JD, Crick FHC. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*. 1953;**171**:737-738. DOI: 10.1038/171737a0
- [3] Sanger F, Nicklen S, Coulson A. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*. 1977;**74**:5463-5467. DOI: 10.1073/PNAS.74.12.5463
- [4] Chidgeavadze ZG, Beabealashvili RS, Atrazhev AM, Kukhanova MK, Azhayev AV, Krayevsky AA. 2',3'-Dideoxy-3' amlnonudeo 5' triphosphates are the terminators of DNA synthesis catalyzed by DNA polymerases. *Nucleic Acids Research*. 1984;**12**:1671-1686. DOI: 10.1093/NAR/12.3.1671
- [5] Smith LM, Fung S, Hunkapiller MW, Hunkapiller TJ, Hood LE. The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: Synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic Acids Research*. 1985;**13**: 2399-2412. DOI: 10.1093/NAR/13.7.2399
- [6] Ansorge W, Sproat BS, Stegemann J, Schwager C. A non-radioactive automated method for DNA sequence determination. *Journal of Biochemical and Biophysical Methods*. 1986;**13**: 315-323. DOI: 10.1016/0165-022X(86) 90038-2
- [7] Ansorge W, Sproat B, Stegemann J, Schwager C, Zenke M. Automated DNA sequencing: Ultrasensitive detection of fluorescent bands during electrophoresis. *Nucleic Acids Research*. 1987;**15**:4593-4602. DOI: 10.1093/NAR/ 15.11.4593
- [8] Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, Cocuzza AJ, Jensen MA, Baumeister K, et al. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science*. 1987;**238**:336-341. DOI: 10.1126/SCIENCE.2443975
- [9] Kambara H, Nishikawa T, Katayama Y, Yamaguchi T. Optimization of parameters in a DNA sequenator using fluorescence detection. *Nature Biotechnology*. 1988; **6**:816-821. DOI: 10.1038/nbt0788-816
- [10] Luckey JA, Drossman H, Kostichka AJ, Mead DA, D'Cunha J, Norris TB, et al. High speed DNA sequencing by capillary electrophoresis. *Nucleic Acids Research*. 1990;**18**: 4417-4421. DOI: 10.1093/NAR/18.15.4417
- [11] Swerdlow H, Gesteland R. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Research*. 1990;**18**:1415-1419. DOI: 10.1093/NAR/18.6.1415
- [12] Jackson DA, Symons RH, Berg P. Biochemical method for inserting new genetic information into DNA of simian virus 40: Circular SV40 DNA molecules containing lambda phage genes and the galactose operon of *Escherichia coli*. *Proceedings of the National Academy of Sciences*. 1972;**69**:2904-2909. DOI: 10.1073/PNAS.69.10.2904
- [13] Cohen SN, Chang ACY, Boyer HW, Helling RB. Construction of biologically functional bacterial plasmids in vitro. *Proceedings of the National Academy of Sciences*. 1973;**70**:3240-3244. DOI: 10.1073/PNAS.70.11.3240
- [14] Nyrén P, Lundin A. Enzymatic method for continuous monitoring of

- inorganic pyrophosphate synthesis. *Analytical Biochemistry*. 1985;**151**: 504-509. DOI: 10.1016/0003-2697(85)90211-8
- [15] Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: From basic research to diagnostics. *Clinical Chemistry*. 2009; **55**:641-658. DOI: 10.1373/CLINCHEM.2008.112789
- [16] Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008;**456**:53-59. DOI: 10.1038/nature07517
- [17] Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Research*. 2006;**34**:e22-e22. DOI: 10.1093/NAR/GNJ023
- [18] McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, et al. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research*. 2009;**19**:1527-1541. DOI: 10.1101/GR.091868.109
- [19] Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*. 2010;**327**: 78-81. DOI: 10.1126/SCIENCE.1181498
- [20] Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*. 2011;**475**:348-352. DOI: 10.1038/nature10242
- [21] Braslavsky I, Hebert B, Kartalov E, Quake SR. Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences*. 2003;**100**:3960-3964. DOI: 10.1073/PNAS.0230489100
- [22] Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, et al. Single-molecule DNA sequencing of a viral genome. *Science*. 2008;**320**:106-109. DOI: 10.1126/SCIENCE.1150427
- [23] Thermes C. Ten years of next-generation sequencing technology. *Trends in Genetics*. 2014;**30**:418-426. DOI: 10.1016/J.TIG.2014.07.001
- [24] Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*. 2003;**299**:682-686. DOI: 10.1126/SCIENCE.1079700
- [25] Haque F, Li J, Wu HC, Liang XJ, Guo P. Solid-state and biological nanopore for real-time sensing of single chemical and sequencing of DNA. *Nano Today*. 2013;**8**:56-74. DOI: 10.1016/J.NANTOD.2012.12.008
- [26] Kasianowicz JJ, Brandin E, Branton D, Deamer DW. Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences*. 1996;**93**: 13770-13773. DOI: 10.1073/PNAS.93.24.13770
- [27] Clarke J, Wu H-C, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*. 2009;**4**:265-270. DOI: 10.1038/nnano.2009.12
- [28] Eisenstein M. Oxford nanopore announcement sets sequencing sector abuzz. *Nature Biotechnology*. 2012;**30**: 295-296. DOI: 10.1038/NBT0412-295
- [29] Loman NJ, Quinlan AR. Poretools: A toolkit for analyzing nanopore sequence

- data. *Bioinformatics*. 2014;**30**: 3399-3401. DOI: 10.1093/BIOINFORMATICS/BTU555
- [30] Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, et al. The potential and challenges of nanopore sequencing. *Nature Biotechnology*. 2008;**26**: 1146-1153. DOI: 10.1038/nbt.1495
- [31] Messina I, Cabras T, Iavarone F, Vincenzoni F, Urbani A, Castagnola M. Unraveling the different proteomic platforms. *Journal of Separation Science*. 2013;**36**:128-139. DOI: 10.1002/JSSC.201200830
- [32] Rochfort S. Metabolomics reviewed: A new “Omics” platform technology for systems biology and implications for natural products research. *Journal of Natural Products*. 2005;**68**:1813-1820. DOI: 10.1021/NP050255W
- [33] Foulston L. Genome mining and prospects for antibiotic discovery. *Current Opinion in Microbiology*. 2019; **51**:1-8. DOI: 10.1016/j.mib.2019.01.001
- [34] Zerikly M, Challis GL. Strategies for the discovery of new natural products by genome mining. *Chembiochem*. 2009;**10**:625-633. DOI: 10.1002/cbic.200800389
- [35] Albarano L, Esposito R, Ruocco N, Costantini M. Genome mining as new challenge in natural products discovery. *Marine Drugs*. 2020;**18**:1-17
- [36] Martin JF, Liras P. Organization and expression of genes involved in the biosynthesis of antibiotics and other secondary metabolites. *Annual Review of Microbiology*. 1989;**43**:173-206. DOI: 10.1146/annurev.mi.43.100189.001133
- [37] Wright L, Hopwood D. Identification of the antibiotic determined by the SCPl. *Journal of General Microbiology*. 1975;**95**:96-106
- [38] Ziemert N, Alanjary M, Weber T. Natural product reports the evolution of genome mining in microbes—A review. *Natural Product Reports*. 2016;**33**: 988-1005. DOI: 10.1039/C6NP00025H
- [39] Donadio S, Staver MJ, McAlpine JB, Swanson SJ, Katz L. Modular organization of genes required for complex polyketide biosynthesis. *Science*. 1991;**205**:675-679
- [40] Beutler B, Hoebe K, Du X, Ulevitch RJ. How we detect microbes and respond to them: The Toll-like receptors and their transducers. *Journal of Leukocyte Biology*. 2003;**74**:479-485. DOI: 10.1189/jlb.0203082
- [41] Bachmann BO, Van Lanen SG, Baltz RH. Microbial genome mining for accelerated natural products discovery: Is a renaissance in the making? *Journal of Industrial Microbiology & Biotechnology*. 2014;**41**:175-184. DOI: 10.1007/s10295-013-1389-9
- [42] Nett M. Genome mining: Concept and strategies for natural product discovery. *Progress in the Chemistry of Organic Natural Products*. 2014;**99**:199-245. DOI: 10.1007/978-3-319-04900-7_4.
- [43] Baltz RH. Synthetic biology, genome mining, and combinatorial biosynthesis of NRPS—Derived antibiotics: A perspective. *Journal of Industrial Microbiology and Biotechnology*. 2018;**45**:635-649. DOI: 10.1007/s10295-017-1999-8
- [44] Sekurova ON, Schneider O, Zotchev SB. Novel bioactive natural products from bacteria via bioprospecting, genome mining and metabolic engineering. *Microbial Biotechnology*. 2019;**12**:828-844. DOI: 10.1111/1751-7915.13398
- [45] Omura S, Ikeda H, Ishikawa J, Hanamoto A, Takahashi C, Shinose M, et al. Genome sequence of an industrial microorganism *Streptomyces avermitilis*:

Deducing the ability of producing secondary metabolites. *PNAS*. 2001;**98**: 12215-12220

[46] Chater KF. Genetics of differentiation in streptomyces. *Annual Review of Microbiology*. 1993;**47**:685-713

[47] Behnken S, Hertweck C. Anaerobic bacteria as producers of antibiotics. *Applied Microbiology and Biotechnology*. 2012;**96**:61-67. DOI: 10.1007/s00253-012-4285-8

[48] Welker M, Dittmann E, Von Döhren H. Cyanobacteria as a source of natural products. *Methods in Enzymology*. 2012;**517**:23-46. DOI: 10.1016/B978-0-12-404634-4.00002-4

[49] Katz M, Hover BM, Brady SF. Culture-independent discovery of natural products from soil metagenomes. *Journal of Industrial Microbiology & Biotechnology*. 2016; **43**:129-141. DOI: 10.1007/s10295-015-1706-6

[50] Zhang X, Wang TT, Xu QL, Xiong Y, Zhang L, Han H, et al. Genome mining and comparative biosynthesis of meroterpenoids from two phylogenetically distinct fungi. *Angewandte Chemie*. 2018;**180**: 8184–8188 DOI: 10.1002/ange.201804317

[51] Medema MH, Cimermancic P, Sali A, Takano E, Fischbach MA. A systematic computational analysis of biosynthetic gene cluster evolution: Lessons for engineering biosynthesis. *PLoS Computational Biology*. 2014;**10**: e1004016. DOI: 10.1371/journal.pcbi.1004016

[52] Hadjithomas M, Chen IA, Chu K, Ratner A, Palaniappan K, Szeto E, et al. IMG-ABC: A knowledge base to fuel discovery of biosynthetic gene clusters and novel secondary metabolites. *MBio*. 2015;**6**:1-10. DOI: 10.1128/mBio.00932-15.Editor

[53] Nogales A, Martínez-sobrido L. Reverse genetics approaches for the development of influenza vaccines. *International Journal of Molecular Sciences*. 2017;**18**:1-26. DOI: 10.3390/ijms18010020

[54] Farnet CM, Zazopoulos E. Improving drug discovery from microorganisms. In: *Natural Products: Drug Discovery and Therapeutic Medicine*. 2005. pp. 95-106. Humana Press (Totowa, US). DOI: 10.1007/978-1-59259-976-9_5

[55] Timmermans ML, Paudel YP, Ross AC. Investigating the biosynthesis of natural products from marine proteobacteria: A survey of molecules and strategies. *Marine Drugs*. 2017;**15**. p. 235 DOI: 10.3390/md15080235

[56] Lee M, Philippe J, Katsanis N, Zhou W. Polyketide synthase plays a conserved role in otolith formation. *Zebrafish*. 2019;**16**:363-369. DOI: 10.1089/zeb.2019.1734

[57] Adhikari K, Lo I, Chen C, Wang Y, Lin K, Zadeh SM, et al. Chemoenzymatic synthesis and biological evaluation for bioactive molecules derived from bacterial benzoyl coenzyme A ligase and plant type III polyketide synthase. *Biomolecules*. 2020;**10**. p. 738

[58] Maansson M, Vynne NG, Klitgaard A, Nybo JL, Melchiorson J, Nguyen DD, et al. An integrated metabolomic and genomic mining workflow to uncover the biosynthetic potential of bacteria. *mSystems*. 2016;**1**: 1-14. DOI: 10.1128/mSystems.00028-15. Editor

[59] Olano C, Méndez C, Salas JA. Strategies for the design and discovery of novel antibiotics using genetic engineering and genome mining. In: *Antimicrobial Compounds*. 2014. pp. 1-25. Springer (Berlin, Germany). DOI: 10.1007/978-3-87-642-40444-3

- [60] Hannigan GD, Prihoda D, Palicka A, Soukup J, Klempir O, Rampula L, et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Research*. 2019;**47**e110. DOI: 10.1093/nar/gkz654
- [61] Russell AH, Truman AW. Genome mining strategies for ribosomally synthesised and post-translationally modified peptides. *Computational and Structural Biotechnology Journal*. 2020; **18**:1838-1851. DOI: 10.1016/j.csbj.2020.06.032
- [62] Su Z, Chen X, Liu X, Guo Q, Li S, Lu X, et al. Genome mining and UHPLC-QTOF-MS/MS to identify the potential antimicrobial compounds and determine the specificity of biosynthetic gene clusters in *Bacillus subtilis* NCD-2. *BMC Genomics*. 2020;**21**:1-16
- [63] Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee Y, et al. antiSMASH 5. 0: Updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Research*. 2019; **47**:81-87. DOI: 10.1093/nar/gkz310
- [64] Hammami R, Zouhir A, Le Lay C, Ben HJ, Fliss I. BACTIBASE second release: A database and tool platform for bacteriocin characterization. *BMC Microbiology*. 2010;**10**:1-5
- [65] Tremblay N, Hill P, Conway KR, Boddy CN. The use of ClusterMine360 for the analysis of polyketide and nonribosomal peptide biosynthetic pathways. *Methods in Molecular Biology*. 2016;1401:233-52. DOI: 10.1007/978-1-4939-3375-4_15
- [66] Diminic J, Zucko J, Trninic I, Cullum J, Starcevic A. Databases of the thiotemplate modular systems (CSDB) and their in silico recombinants (r-CSDB). *Journal of Industrial Microbiology and Biotechnology*. 2013; **40**:653-659. DOI: 10.1007/s10295-013-1252-z
- [67] Ichikawa N, Sasagawa M, Yamamoto M, Komaki H, Yoshida Y. DoBISCUIT: A database of secondary metabolite biosynthetic gene clusters. *Nucleic Acids Research*. 2013;**41**:408-414. DOI: 10.1093/nar/gks1177
- [68] Palaniappan K, Chen IA, Chu K, Ratner A, Seshadri R, Kyrpides NC, et al. IMG-ABC v.5.0: An update to the IMG/atlas of biosynthetic gene clusters knowledgebase. *Nucleic Acids Research*. 2019;**48**:422-430. DOI: 10.1093/nar/gkz932
- [69] Alanjary M, Blin K, Weber T, Medema MH, Ziemert N. ARTS 2.0: Feature updates and expansion of the Antibiotic Resistant Target Seeker for comparative genome mining. *Nucleic Acids Research*. 2020;**48**:546-552. DOI: 10.1093/nar/gkaa374
- [70] Navarro-Muñoz JC, Selem-mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, et al. A computational framework to explore large-scale biosynthetic diversity. *Nature Chemical Biology*. 2020;**16**:60-68. DOI: 10.1038/s41589-019-0400-9
- [71] Weber T, Rausch C, Lopez P, Hoof I, Gaykova V, Huson DH, et al. CLUSEAN: A computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. *Journal of Biotechnology*. 2009;**140**:13-17. DOI: 10.1016/j.jbiotec.2009.01.007
- [72] Cimermanic P, Medema MH, Claesen J, Kurita K, Brown LCW, Mavrommatis K, et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell*. 2010;**158**:412-421. DOI: 10.1016/j.cell.2014.06.034
- [73] Cruz-Morales P, Kopp JF, Martí C, Barona-go F, Selem-mojica N, Ramos-aboites H. Phylogenomic analysis of natural products biosynthetic gene

- p>clusters allows discovery of arseno-organic metabolites in model streptomycetes.
- Genome Biology and Evolution*
- . 2016;
- 8**
- :1906-1916. DOI: 10.1093/gbe/evw125
- [74] Andersen MR, Nielsen JB, Klitgaard A, Petersen LM, Zachariasen M, Hansen TJ. Accurate prediction of secondary metabolite gene clusters in filamentous fungi. *PNAS*. 2012;**110**: 100-107. DOI: 10.1073/pnas.1205532110
- [75] Umemura M, Koike H, Machida M. Motif-independent de novo detection of secondary metabolite gene clusters—Toward identification from filamentous fungi. *Frontiers in Microbiology*. 2015;**6**: 1-14. DOI: 10.3389/fmicb.2015.00371
- [76] Ziemert N, Podell S, Penn K, Badger JH, Allen E, Jensen PR. The natural product domain seeker NaPDoS: A phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLoS One*. 2012;**7**:1-9. DOI: 10.1371/journal.pone.0034064
- [77] Nadine T, Fuchs L, Aharoni A. The PhytoClust tool for metabolic gene clusters discovery in plant genomes. *Nucleic Acids Research*. 2017;**45**: 7049-7063. DOI: 10.1093/nar/gkx404
- [78] Kim J, Yi G. PKMiner: A database for exploring type II polyketide synthases. *BMC Microbiology*. 2012;**12**: 1-12
- [79] Blin K, Andreu P, Santos ELCDL, Del Carratore F, Lee SY, Medema MH, et al. The antiSMASH database version 2: A comprehensive resource on secondary metabolite biosynthetic gene. *Nucleic Acids Research*. 2019;**47**: 625-630. DOI: 10.1093/nar/gky1060
- [80] Staniek A, Bouwmeester H, Fraser PD, Kayser O, Martens S, Tissier A, et al. Natural products—Modifying metabolite pathways in plants. *Biotechnology Journal*. 2013;**8**:1159-1171. DOI: 10.1002/BIOT.201300224
- [81] Staniek A, Bouwmeester H, Fraser PD, Kayser O, Martens S, Tissier A, et al. Natural products—Learning chemistry from plants. *Biotechnology Journal*. 2014;**9**:326-336. DOI: 10.1002/BIOT.201300059
- [82] Kautsar SA, Duran HGS, Blin K, Osbourn A, Medema H. plantiSMASH: Automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Research*. 2017;**45**:55-63. DOI: 10.1093/nar/gkx305
- [83] Khater S, Gupta M, Agrawal P, Sain N, Prava J, Gupta P, et al. SBSPKSV2: Structure-based sequence analysis of polyketide synthases and non-ribosomal peptide synthetases. *Nucleic Acids Research*. 2017;**45**:72-79. DOI: 10.1093/nar/gkx344
- [84] Khaldi N, Seifuddin FT, Turner G, Haft D, Nierman WC, Wolfe KH, et al. SMURF: Genomic mapping of fungal secondary metabolite clusters. *Fungal Genetics and Biology*. 2010;**47**: 736-741. DOI: 10.1016/j.fgb.2010.06.003
- [85] Leblond P, Gondry M, Juguet M, Lautru S. An iterative nonribosomal peptide synthetase assembles the pyrrole-amide antibiotic congocidine in *Streptomyces ambofaciens*. *Cell Chemical Biology*. 2009;**2820**:421-431. DOI: 10.1016/j.chembiol.2009.03.010
- [86] Kang HS. Phylogeny—Guided (meta) genome mining approach for the targeted discovery of new microbial natural products. *Journal of Industrial Microbiology & Biotechnology*. 2017;**44**:285-293. DOI: 10.1007/s10295-016-1874-z
- [87] Farag MA, Weigend M, Luebert F, Brokamp G, Wessjohann LA. Phytochemical, phylogenetic, and anti-inflammatory evaluation of 43 *Urtica* accessions (stinging nettle) based on UPLC-Q-TOF-MS metabolomic

- profiles. *Phytochemistry*. 2013;**96**: 170-183. DOI: 10.1016/J.PHYTOCHEM.2013.09.016
- [88] Abdelmohsen UR, Yang C, Horn H, Hajjar D, Ravasi T, Hentschel U. Actinomycetes from red sea sponges: Sources for chemical and phylogenetic diversity. *Marine Drugs*. 2014;**12**: 2771-2789. DOI: 10.3390/md12052771
- [89] Singh SP, Klisch M, Sinha RP, Häder D. Genome mining of mycosporine-like amino acid (MAA) synthesizing and non-synthesizing cyanobacteria: A bioinformatics study. *Genomics*. 2010;**95**:120-128. DOI: 10.1016/j.ygeno.2009.10.002
- [90] Eisen JA, Wu M. Phylogenetic analysis and gene functional predictions: Phylogenomics in action. *Theoretical Population Biology*. 2002;**61**:481-487. DOI: 10.1006/tpbi.2002.1594
- [91] Ward AC, Allenby NEE. Genome mining for the search and discovery of bioactive compounds: The Streptomyces paradigm. *FEMS Microbiology Letters*. 2018;**365**:1-20. DOI: 10.1093/femsle/fny240
- [92] Labreuche Y, Krin E, Ansquer D, Goudene D, Mangenot S, Calteau A, et al. Comparative genomics of pathogenic lineages of *Vibrio nigrapulchritudo* identifies virulence-associated traits. *The ISME Journal*. 2013;**93**:1985-1996. DOI: 10.1038/ismej.2013.90
- [93] Cuadrat RRC, Ionescu D, Dávila AMR, Marco DE. Recovering Genomics clusters of secondary metabolites from lakes using genome-resolved metagenomics. *Frontiers in Microbiology*. 2018;**9**:1-13. DOI: 10.3389/fmicb.2018.00251
- [94] Kurita KL, Glassey E, Linington RG. Integration of high-content screening and untargeted metabolomics for comprehensive functional annotation of natural product libraries. *PNAS*. 2015; **112** p. 1 to 6. DOI: 10.1073/pnas.1507743112
- [95] Bhushan A, Egli PJ, Peters EE, Freeman MF, Piel J. Genome mining- and synthetic biology-enabled production of hypermodified peptides. *Nature Chemistry*. 2019;**11**:931-939. DOI: 10.1038/s41557-019-0323-9
- [96] Machado H, Tuttle RN, Jensen PR. Omics-based natural product discovery and the lexicon of genome mining. *Current Opinion in Microbiology*. 2017; **39**:136-142. DOI: 10.1016/j.mib.2017.10.025
- [97] Moreno-Pachon NM, Leeggangers HACF, Nijveen H, Severing E, Hilhorst H, Immink RGH. Elucidating and mining the Tulipa and Lilium transcriptomes. *Plant Molecular Biology*. 2016;**92**:249-261. DOI: 10.1007/S11103-016-0508-1
- [98] Nützmann H-W, Huang A, Osbourn A. Plant metabolic clusters—From genetics to genomics. *The New Phytologist*. 2016;**211**:771-789. DOI: 10.1111/NPH.13981
- [99] Jiang C, Zhang H, Ren J, Dong J, Zhao X, Wang X, et al. Comparative transcriptome-based mining and expression profiling of transcription factors related to cold tolerance in peanut. *International Journal of Molecular Sciences*. 2020;**21**:1921. DOI: 10.3390/IJMS21061921
- [100] Yadav R, Verma OP, Padaria JC. Transcript profiling and gene expression analysis under drought stress in *Ziziphus nummularia* (Burm.f.) Wright & Arn. *Molecular Biology Reports*. 2018;**45**: 163-174. DOI: 10.1007/S11033-018-4149-0
- [101] Nath VS, Koyyappurath S, Alex TE, Geetha KA, Augustine L, Nasser A, et al. Transcriptome-based mining and expression profiling of Pythium responsive transcription factors in Zingiber sp. *Functional &*

- Integrative Genomics. 2018;**19**:249-264. DOI: 10.1007/S10142-018-0644-6
- [102] Sharma G, Aminedi R, Saxena D, Gupta A, Banerjee P, Jain D, et al. Effector mining from the *Erysiphe pisi* haustorial transcriptome identifies novel candidates involved in pea powdery mildew pathogenesis. Molecular Plant Pathology. 2019;**20**:1506-1522. DOI: 10.1111/MPP.12862
- [103] Jongeneel CV. Searching the expressed sequence tag (EST) databases: Panning for genes. Briefings in Bioinformatics. 2000;**1**:76-92. DOI: 10.1093/BIB/1.1.76
- [104] Lamblin A-FJ, Crow JA, Johnson JE, Silverstein KAT, Kunau TM, Kilian A, et al. MtDB: A database for personalized data mining of the model legume *Medicago truncatula* transcriptome. Nucleic Acids Research. 2003;**31**:196-201. DOI: 10.1093/NAR/GKG119
- [105] Kelso J, Visagie J, Theiler G, Christoffels A, Bardien S, Smedley D, et al. eVOC: A controlled vocabulary for unifying gene expression data. Genome Research. 2003;**13**:1222-1230. DOI: 10.1101/GR.985203
- [106] Yonekura-Sakakibara K, Fukushima A, Saito K. Transcriptome data modeling for targeted plant metabolic engineering. Current Opinion in Biotechnology. 2013;**24**:285-290. DOI: 10.1016/J.COPBIO.2012.10.018
- [107] Rao X, Dixon RD. Co-expression networks for plant biology: Why and how. Acta Biochimica et Biophysica Sinica. 2019;**51**:981-988. DOI: 10.1093/ABBS/GMZ080
- [108] Fang X, Li C-Y, Yang Y, Cui M-Y, Chen X-Y, Yang L. Identification of a novel (–)-5-epieremophilene synthase from *Salvia miltiorrhiza* via transcriptome mining. Frontiers in Plant Science. 2017;**0**:627. DOI: 10.3389/FPLS.2017.00627
- [109] Karunanithi PS, Dhanota P, Addison JB, Tong S, Fiehn O, Zerbe P. Functional characterization of the cytochrome P450 monooxygenase CYP71AU87 indicates a role in marrubiin biosynthesis in the medicinal plant *Marrubium vulgare*. BMC Plant Biology. 2019;**19**:1-14. DOI: 10.1186/S12870-019-1702-5
- [110] Keeling CI, Weisshaar S, Ralph SG, Jancsik S, Hamberger B, Dullat HK, et al. Transcriptome mining, functional characterization, and phylogeny of a large terpene synthase gene family in spruce (*Picea* spp.). BMC Plant Biology. 2011;**11**:1-14. DOI: 10.1186/1471-2229-11-43
- [111] Aminfar Z, Rabiei B, Tohidfar M, Mirjalili MH. Identification of key genes involved in the biosynthesis of triterpenic acids in the mint family. Scientific Reports. 2019;**9**:1-15. DOI: 10.1038/s41598-019-52090-z
- [112] Hodgson H, La Peña RD, Stephenson MJ, Thimmappa R, Vincent JL, Sattely ES, et al. Identification of key enzymes responsible for protolimonoid biosynthesis in plants: Opening the door to azadirachtin production. Proceedings of the National Academy of Sciences. 2019;**116**:17096-17104. DOI: 10.1073/PNAS.1906083116
- [113] Singh A, Massicotte M-A, Garand A, Tousignant L, Ouellette V, Bérubé G, et al. Cloning and characterization of norbelladine synthase catalyzing the first committed reaction in Amaryllidaceae alkaloid biosynthesis. BioMed Central Plant Biol. 2018;**18**:1-12. DOI: 10.1186/S12870-018-1570-4
- [114] Hagel JM, Morris JS, Lee E-J, Desgagné-Penix I, Bross CD, Chang L, et al. Transcriptome analysis of 20 taxonomically related benzyloquinoline alkaloid-producing plants. BMC Plant Biology. 2015;**15**:1-16. DOI: 10.1186/S12870-015-0596-0

- [115] Lau W, Sattely ES. Six enzymes from mayapple that complete the biosynthetic pathway to the etoposide aglycone. *Science*. 2015;**349**: 1224-1228. DOI: 10.1126/SCIENCE.AAC7202
- [116] Rizzo P, Altschmied L, Stark P, Rutten T, Gündel A, Scharfenberg S, et al. Discovery of key regulators of dark gland development and hypericin biosynthesis in St. John's wort (*Hypericum perforatum*). *Plant Biotechnology Journal*. 2019;**17**: 2299-2312. DOI: 10.1111/PBI.13141
- [117] Pires ÁS, Rigueiras PO, Dohms SM, Porto WF, Franco OL. Structure-guided identification of antimicrobial peptides in the spathe transcriptome of the non-model plant, arum lily (*Zantedeschia aethiopica*). *Chemical Biology & Drug Design*. 2019;**93**:1265-1275. DOI: 10.1111/CBDD.13498
- [118] Guerriero G, Piasecki E, Berni R, Xu X, Legay S, Hausman J-F. Identification of callose synthases in stinging nettle and analysis of their expression in different tissues. *International Journal of Molecular Sciences*. 2020;**21**:3853. DOI: 10.3390/IJMS21113853
- [119] Prasad P, Sreedhar RV. Identification and functional characterization of *Buglossoides arvensis* microsomal fatty acid desaturation pathway genes involved in polyunsaturated fatty acid synthesis in seeds. *Journal of Biotechnology*. 2019; **308**:130-140. DOI: 10.1016/J.JBIOTEC.2019.12.006
- [120] Hellinger R, Koehbach J, Soltis DE, Carpenter EJ, Wong GK-S, Gruber CW. Peptidomics of circular cysteine-rich plant peptides: Analysis of the diversity of cyclotides from viola tricolor by transcriptome and proteome mining. *Journal of Proteome Research*. 2015;**14**: 4851-4862. DOI: 10.1021/ACS.JPROTEOME.5B00681
- [121] Ran X, Zhao F, Wang Y, Liu J, Zhuang Y, Ye L, et al. Plant regulomics: A data-driven interface for retrieving upstream regulators from plant multi-omics data. *The Plant Journal*. 2020;**101**: 237-248. DOI: 10.1111/TPJ.14526
- [122] Werner C, Doenst T, Schwarzer M. Metabolic pathways and cycles. In: *The Scientist's Guide to Cardiac Metabolism*. Amsterdam, The Netherlands: Elsevier Inc.; 2016. pp. 39-55. DOI: 10.1016/B978-0-12-802394-5/00004-2
- [123] Czarnecki JM, Shepherd AJ. Metabolic pathway mining. *Methods in Molecular Biology*. 2017;**1526**:139-158. DOI: 10.1007/978-1-4939-6613-4_8
- [124] Rivera R, Garrido N. Metabolomics. In: *Oxidants, Antioxidants, and Impact of the Oxidative Status in Male Reproduction*. Amsterdam, Netherlands: Elsevier. 2018. pp. 277-285. DOI: 10.1016/B978-0-12-812501-4.00025-0
- [125] Carbonell P. Enzyme discovery and selection. *Metabolic Pathway Design: A Practical Guide*. Berlin, Germany: Springer. 2019. pp. 63-81. DOI: 10.1007/978-3-030-29865-4_5
- [126] Giraudeau P. NMR-based metabolomics and fluxomics: Developments and future prospects. *Analyst*. 2020;**145**:2457-2472. DOI: 10.1039/D0AN00142B
- [127] Salon C, Avice JC, Colombié S, Dieuaide-Noubhani M, Gallardo K, Jeudy C, et al. Fluxomics links cellular functional analyses to whole-plant phenotyping. *Journal of Experimental Botany*. 2017;**68**:2083-2098. DOI: 10.1093/JXB/ERX126
- [128] Singh S, Tiwari BS. Biosynthesis of high-value amino acids by synthetic biology. In: *Current Developments in Biotechnology and Bioengineering: Synthetic Biology, Cell Engineering and*

Bioprocessing Technologies.

Amsterdam, Netherlands: DOI: 10.1016/B978-0-444-64085-7.00011-3

[129] Tian H, Zada B, Singh BH, Wang C, Kim SW. Synthetic biology approaches for the production of isoprenoids in *Escherichia coli*. In: Current Developments in Biotechnology and Bioengineering: Synthetic Biology, Cell Engineering and Bioprocessing Technologies. Amsterdam, Netherlands: Elsevier; 2018. Elsevier (Amsterdam, Netherlands). DOI: 10.1016/B978-0-444-64085-7.00013-7

[130] Saini DK, Pabbi S, Prakash A, Shukla P. Synthetic biology applied to microalgae-based processes and products. In: Handbook of Microalgae-Based Processes and Products. Amsterdam, The Netherlands: Elsevier Inc.; 2020. DOI: 10.1016/B978-0-12-818536-0.00004-x

[131] Feiner A, Pitra N, Matthews P, Pillen K, Wessjohann LA, Riewe D. Downy mildew resistance is genetically mediated by prophylactic production of phenylpropanoids in hop. *Plant, Cell & Environment*. 2021;44:323-338. DOI: 10.1111/PCE.13906

[132] Ranganathan S, Zhao Y, Simon R. Encyclopedia of Systems Biology. Berlin, Germany: Springer; 2013. DOI: 10.1007/978-1-4419-9863-7

[133] Vick JE, Johnson ET, Choudhary S, Bloch SE, Lopez-Gallego F, Srivastava P, et al. Optimized compatible set of BioBrick™ vectors for metabolic pathway engineering. *Applied Microbiology and Biotechnology*. 2011; 92:1275-1286. DOI: 10.1007/s00253-011-3633-4

[134] Camp D, Garavelas A, Campitelli M. Analysis of Physicochemical Properties for Drugs of Natural Origin. *Journal of Natural Products*. 2015;78:1370-1382. DOI: 10.1021/acs.jnatprod.5b00255

[135] Haustedt LO, Siems K. The role of natural products in drug discovery: Examples of marketed drugs. In: Small Molecule Medicinal Chemistry: Strategies and Technologies. Hoboken, US: John Wiley & Sons, Inc.; 2016. pp. 381-430.

[136] Newman DJ, Cragg GM. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *Journal of Natural Products*. 2020;83:770-803. DOI: 10.1021/ACS.JNATPROD.9B01285

[137] Wessjohann LA. Reverse metabolomics—Metabolomics in drug discovery: Connecting metabolomic profiles with phylogenetic, medicinal and flavoring properties. *Metabolomics*. 2014;s1. DOI: 10.4172/2153-0769.S1.024

[138] Zhou JZ. Chemoinformatics and library design. In: Chemical Library Design: Methods and Protocols. Berlin, Germany: Springer; 2011. pp. 27-52. Springer (Berlin, Germany). DOI: 10.1007/978-1-60761-931-4

[139] Santana K, do Nascimento LD, Lima e Lima A, Damasceno V, Nahum C, Braga RC, et al. Applications of virtual screening in bioprospecting: Facts, shifts, and perspectives to explore the chemo-structural diversity of natural products. *Frontiers in Chemistry*. 2021;9. Article: 662688. DOI: 10.3389/fchem.2021.662688

[140] Reymond J-L, Awale M. Exploring chemical space for drug discovery using the chemical universe database. *ACS Chemical Neuroscience*. 2012;3:649-657. DOI: 10.1021/cn3000422

[141] Liu R, Li X, Lam KS. Combinatorial chemistry in drug discovery. *Current Opinion in Chemical Biology*. 2017;38: 117-126. DOI: 10.1016/j.cbpa.2017.03.017

[142] Wessjohann LA. Synthesis of natural-product-based compound

libraries. *Current Opinion in Chemical Biology*. 2000;**4**:303-309. DOI: 10.1016/S1367-5931(00)00093-4

helpful strategy to promote a fragment “Rescue” effect. *Frontiers in Chemistry*. 2019;**7**:1-7. DOI: 10.3389/fchem.2019.00564

[143] Kodadek T. The rise, fall and reinvention of combinatorial chemistry. *Chemical Communications*. 2011;**47**: 9757-9763. DOI: 10.1039/c1cc12102b

[144] Schulz MN, Hubbard RE. Recent progress in fragment-based lead discovery. *Current Opinion in Pharmacology*. 2009;**9**:615-621. DOI: 10.1016/j.coph.2009.04.009

[145] Erlanson DA, Fesik SW, Hubbard RE, Jahnke W, Jhoti H. Twenty years on: The impact of fragments on drug discovery. *Nature Reviews. Drug Discovery*. 2016;**15**: 605-619. DOI: 10.1038/nrd.2016.109

[146] Chen H, Zhou X, Wang A, Zheng Y, Gao Y, Zhou J. Evolutions in fragment-based drug design: The deconstruction-reconstruction approach. *Drug Discovery Today*. 2015; **20**:105-113. DOI: 10.1016/j.drudis.2014.09.015

[147] Ahmed J, Worth CL, Thaben P, Matzig C, Blasse C, Dunkel M, et al. FragmentStore—A comprehensive database of fragments linking metabolites, toxic molecules and drugs. *Nucleic Acids Research*. 2011;**39**: 1049-1054. DOI: 10.1093/nar/gkq969

[148] Scott DE, Coyne AG, Hudson SA, Abell C. Fragment based approaches in drug discovery and chemical biology. *Biochemistry*. 2012;**51**:4990-5003. DOI: 10.1021/bi3005126

[149] Keserü GM, Makara GM. The influence of lead discovery strategies on the properties of drug candidates. *Nature Reviews. Drug Discovery*. 2009; **8**:203-212. DOI: 10.1038/nrd2796

[150] Vásquez AF, González BA. Pushing the ligand efficiency metrics: Relative group contribution (RGC) model as a