BiosyntheticSPAdes: Reconstructing Biosynthetic Gene Clusters From Assembly Graphs

Dmitry Meleshko, Hosein Mohimani, Vittorio Traccana, et al.

Genome Res. published online June 3, 2019
Access the most recent version at doi:10.1101/gr.243477.118

BiosyntheticSPAdes: Reconstructing Biosynthetic Gene Clusters From Assembly Graphs

Dmitry Meleshko[1,2], Hosein Mohimani[3,4], Vittorio Tracanna[5], Iman Hajirasouliha[6,7], Marnix H. Medema[5], Anton Korobeynikov[1,8], Pavel A. Pevzner[1,3,*]

[1]Center for Algorithmic Biotechnology, Institute for Translational Biomedicine, St. Petersburg State University, St. Petersburg, Russia

[2]Tri-Institutional PhD Program in Computational Biology and Medicine, Weill Cornell Medical College, New York, United States

[3]Department of Computer Science and Engineering, University of California, San Diego,

[4]Computational Biology Department, School of Computer Sciences, Carnegie Mellon University

[5]Bioinformatics Group, Wageningen University, Wageningen, The Netherlands

[6]Institute for Computational Biomedicine, Department of Physiology and Biophysics, Weill Cornell Medicine of Cornell University, New York, United States

[7]Englander Institute for Precision Medicine, Meyer Cancer Center, Weill Cornell Medicine, New York, United States

[8]Department of Statistical Modelling, St. Petersburg State University, St. Petersburg, Russia

*Corresponding author, ppevzner@ucsd.edu

ABSTRACT

Predicting Biosynthetic Gene Clusters (BGCs) is critically important for discovery of antibiotics and other natural products. While BGC prediction from complete genomes is a well-studied problem, predicting BGC in fragmented genomic assemblies remains challenging. The existing BGC prediction tools often assume that each BGC is encoded within a *single* contig in the genome assembly, a condition

that is violated for most sequenced microbial genomes where BGCs are often scattered through several contigs, making it difficult to reconstruct them. The situation is even more severe in shotgun metagenomics, where the contigs are often short, and the existing tools fail to predict a large fraction of long BGCs. While it is difficult to assemble BGCs in a single contig, the structure of the genome assembly graph often provides clues on how to combine multiple contigs into segments encoding long BGCs. We describe biosyntheticSPAdes, a tool for predicting BGCs in assembly graphs and demonstrate that it greatly improves the reconstruction of BGCs from genomic and metagenomics datasets.

INTRODUCTION

Although there exist many tools for assembling microbial genomes or metagenomes (Simpson et al. 2009, Li et al. 2015, Nurk et al. 2017), they all have limitations with respect to assembling contigs that contain long genes encoding proteins with repetitive domains. Since long genes are often scattered between multiple contigs in fragmented assemblies, the existing gene prediction tools (Besemer et al. 2005, Delcher et al. 2007, Pati et al. 2010, Hyatt et al. 2010) cannot predict them. The challenge of assembling long genes in a single contig is illustrated by genes encoding *Nonribosomal Peptides Synthetases* (*NRPSs*), *Polyketide Synthases (PKSs),* and other genes that are parts of *biosynthetic gene clusters (BGCs)* encoding the production of antibiotics and other natural products. BGCs usually include multiple consecutive genes that participate in a single metabolic pathway responsible for synthesizing a natural product. NRPS BGCs encode *Nonribosomal Peptides* (*NRPs*) built from amino acids and PKS BGCs encode *polyketides* (*PSs*) built from keto groups. Mixed NRPS/PKS BGCs contain both NRPS-specific and PKS-specific domains and their natural products represent fusions of peptides and polyketides (Cane et al. 1999). Klassen and Currie, 2012 showed that long and repetitive NRPSs and PKSs are responsible for a large fraction of fragmentation in microbial assemblies.

This paper focuses on NRPSs because NRPs represent an important class of natural product drugs (Newman and Cragg, 2016) that is most amenable to downstream peptidogenomics analysis as compared to other classes of natural products (Kersten et al., 2011, Mohimani et al., 2014, Medema et al., 2014). NRPS BGCs constitute 34% of all BGCs in publicly available genomes, as found in the antiSMASH database (https://antismash-db.secondarymetabolites.org/#!/stats). Since NRPSs are very common (albeit elusive) in diverse bacterial datasets (Mukherjee et al., 2017) and since the downstream peptidogenomics analysis of NRPs is greatly impaired by fragmented assemblies, most examples in this paper refers to NRPs. In addition to NRPS BGCs, biosynteticSPAdes is also applicable to PKS BGCs and mixed NRPS-PKS BGCs (NRPS, PKS, and mixed NRPS-PKS BGCs constitute the majority of BGCs in the MIBiG database). Klassen and Currie, 2012 have shown that fragmented ORFs in genome assemblies are highly enriched in NRPSs and PKSs, which thus constitute a prominent source of breakpoint in (meta)genome assemblies. The fact that the vast majority of genomes contain either an NRPS or a PKS, or a mixed NRPS-PKS BGC (for some species, over 30% of the genome is allocated to these BGCs) and direct interest to a large research community is a good reason to provide a specialized assembler for these BGCs.

NRPSs are large modular protein complexes containing multiple highly similar *adenylation domains* (*A-domains)* responsible for recruiting amino acids that form NRPs according to the substrate specificity of each A-domain (Stachelhaus and Marahiel 1999). NRPSs are often accompanied by other adjacently located genes that together form NRP BGCs and contribute to NRP synthesis, transport, and regulation. NRP BGCs are typically long with an average length of ~60 kb and some exceeding 100 kb in length. Assembling NRP BGCs into single contigs is a crucial step in natural product discovery by genome mining (Weber et al. 2015) and peptidogenomics (Mohimani et al., 2014, Mohimani and Pevzner, 2016, Mohimani et al., 2017, Gurevich et al., 2018).

The recent *Genomic Encyclopedia of Bacteria and Archaea (GEBA)* study of over 1000 bacterial

genomes revealed over 23,000 BGCs (Mukherjee et al., 2017). An average GEBA genome devotes nearly 10% of its genome to BGCs (some genomes devote >30%). However, the vast majority of predicted BCG products remain unknown, in part due to difficulties in predicting long BGCs (Hadjithomas et al., 2015).

The recently proposed genome mining and peptidogenomic approaches elucidate the amino acid sequences of NRPs by matching tandem mass spectra against predicted NRP synthetases in the assembled genomes (Mohimani et al. 2014, Medema et al., 2014, Mohimani et al. 2017). The success of these approaches depends on accurate prediction of genes encoding NRP synthetases followed by machine-learning algorithms to predict their substrate specificities, and matching mass spectral datasets against the predicted NRP amino acid sequences. This is a challenging task requiring the recovery of the *complete* NRPS genes and the corresponding NRP BGCs in a single contig.

This challenge is further amplified in metagenomics assemblies, because NRP synthetases from different species within a microbial community often share similar domains. This makes it difficult to assemble them in a single contig in cases when multiple domains are collapsed into a single edge in the assembly graph (Coates et al. 2014). Therefore, while metagenomes represent a gold mine for antibiotics discovery, a limited number of antibiotics have been discovered from metagenomics datasets so far (Freeman et al. 2012, Donia et. al., 2014, Donia and Fischbach, 2015).

Despite the fact that it is difficult to reconstruct long NRPS BGCs from individual contigs, the structure of the assembly graph often provides clues on how to combine various contigs into intact BGCs. We describe the biosyntheticSPAdes tool for assembling NRPS BGCs in assembly graphs constructed by SPAdes (Bankevich et al., 2012) and metaSPAdes (Nurk et al., 2017) assemblers. Below we show how biosyntheticSPAdes contributes to the discovery of NRPS BGCs in various genomes and metagenomes.

RESULTS

The challenge of assembling BGCs. Contrary to the standard practice in existing gene prediction tools that attempt to reconstruct genes from *individual* contigs/scaffolds, biosyntheticSPAdes analyzes the assembly graph to join fragments of long BGCs (scattered over multiple contigs) into a single contig. Below, we describe the biosyntheticSPAdes algorithm and illustrate how it works using the genome of *Streptomyces coelicolor A3(2)* (referred to as *S. coelicolor* for brevity), a well-studied antibiotics-producing bacterium, which encodes four NRP BGCs (Bentley et al. 2002), including *calcium-dependent antibiotic (CALC)*.

We illustrate the challenge of assembling long repetitive genes using a subgraph of the *S. coelicolor* assembly graph encoding the CALC BGC (Figure 1). To generate this graph, we simulated error-free short paired-end reads (Huang et al. 2012) from the *S. coelicolor* genome using the ART read simulator. The reads from the resulting dataset with coverage 180× (referred to as the STREP dataset and containing paired reads of length 150 bp with mean insert size 300 bp) were assembled using the SPAdes assembler (Bankevich et al. 2012). The assembly graph constructed from these simulated reads contains 626 vertices and 697 edges (484 of them are longer than 1000 bp). The total edge length in the assembly graph is 8,598,860 with N50=41 kb. SPAdes uses paired reads to resolve repeats in the genome and combines some edges in the assembly graphs into contigs/scaffolds using exSPAnder (Prjibelsky et al., 2014). exSPAnder constructed 145 scaffolds longer than 1000 bp with N50=135 kb after the repeat resolution step.

AntiSMASH (Weber et al. 2015) is a popular genome mining tool for detecting and annotating BGCs. AntiSMASH revealed 29 BGCs in the *S. coelicolor* genome, including four NRP BGCs. The CALC BGC with eleven A-domains traverses 25 edges in the assembly graph. exSPAnder (Prjibelsky et al., 2014) combined some of these edges into single contigs, but even after applying exSPAnder, CALC was split

into 7 scaffolds (Figure 1). This illustrates the challenge of reconstructing long genes even for isolated bacteria, let alone metagenomes. Note that 11 A-domains in CALC are represented by only 9 A-domains in Figure 1 because 3 out of 11 A-domains got collapsed into a single edge in the assembly graph.

The CALC BGC illustrates just one example of the difficulties with assembling long and repetitive genes in genomic and metagenomic datasets. Supplementary Table S1 illustrates that 285 out of 7,910 genes (≈ 3%) in the *S. coelicolor* genome are split over multiple edges in the assembly graph. The fraction of split genes further increases when we consider long genes: 11 out of the 100 longest genes (length > 3200 bp) traverse multiple edges and 17 out of these 100 longest genes corresponds to BGCs (Supplementary Table S2). While the repeat resolution step in SPAdes (Prjibelsky et al., 2014) captures some of the split genes in a single contig/scaffold, many long genes remain split even after repeat resolution and three of them correspond to BGC genes (Supplementary Table S3). The fraction of such split genes further increases in metagenomics assemblies.

BiosyntheticSPAdes outline. The biosyntheticSPAdes pipeline includes six steps (Figure 2) that are described in the Methods section:

- assembling genomic/metagenomic reads with SPAdes/metaSPAdes,
- identifying domain-edges in the assembly graph,
- extracting BGC subgraphs from the assembly graph,
- restoring collapsed domains in the assembly graph,
- constructing the scaffolding graph,
- constructing putative BGCs by solving the Rural Postman Problem in the scaffolding graph.

Benchmarking design. To benchmark biosyntheticSPAdes, we compared its output (a single or multiple contigs) against the reference genome(s). Since the downstream applications, such as NRPquest

(Mohimani et al. 2014), do not require a single contig output and work equally well when a small set of output contigs contain a correct one, we classify the biosyntheticSPAdes output as correct if at least one of the reported contigs is contained in one of the reference genomes (with percent identity exceeding 95%).

In the case when the reference genomes are not available, we check whether a BGC subgraph contains a rural postman path. If it is the case, it is likely that one of the reported contigs is contained in an unknown reference genome.

Datasets. We analyzed the following datasets assembled using SPAdes or metaSPAdes with *k*-mer sizes varying from 21 to 55 nucleotides during the iterative assembly.

Pseudomonas datasets (PSEUDO). The PSEUDO dataset (accession number ERR1890333) contains ≈4.5 million paired reads from the isolate of *Pseudomonas protegens (fluorescens) pf-5* (read length 100 bp, a mean insert size 440 bp, and a standard deviation of the insert size 140 bp). The genome sequence was finished using a combination of primer walking, generation and sequencing of transposon-tagged libraries, and multiplex PCR (Paulsen et al. 2005).

Cyanobacteria dataset (CYANO). The CYANO dataset contains genomic reads from cultured marine bacteria *Moorea producens JHB* (referred as JHB below) described in Kleigrewe et al., 2015. The sample is contaminated with heterotrophic bacteria and thus represents a low-complexity metagenome. The JHB strain encodes various NRPs, PKs and mixed NRP-PKs, including *hectochlorin* (Marquez et al. 2002) and *jamaicamides* (Edwards et al. 2004). The JHB dataset contains ≈6 million paired reads (length 150 bp, a mean insert size 292 bp, and a standard deviation of the insert size 74 bp).

MIBiG datasets (MIBIG). The Minimum Information about a Biosynthetic Gene Cluster (MIBiG)

database contains information about BGCs and their products (Medema et al. 2015). Each entry in the MIBiG database contains the nucleotide sequence of a BGC, the natural product type (NRPs, PKs, and other types), and its annotation. In order to benchmark biosyntheticSPAdes on a wide range of BGCs, we extracted all MIBiG entries describing NRPSs and PKSs with complete BGC sequences (665 entries) and used the ART read simulator (Huang et al. 2012) to simulate reads from BGC sequences with the default MiSeq parameters. Admittedly, generating reads from BGCs results in a simpler problem than simulating reads from the entire genome. However, since entire genomes are not available for many MIBiG entries, we simulated reads from BGCs only. We define the *complexity* of a BGC as the total number of A-domains and AT-domains in this BGC. Note that this is a very naïve definition of complexity (e.g., trans-AT PKSs have few AT domains). 139 out of 665 BGCs in the MIBiG dataset have complexity 10 and larger.

HMP datasets (HMP). The HMP dataset consists of 20 metagenomic sub-datasets from seven parts of human body that included keratinized gingiva, buccal mucosa, stool, gingivival plaque, subpravingal plaque, tongue dorsum, and throat (Table S4). The description of these datasets is given in Methé et al. 2012.

Analyzing the PSEUDO dataset. AntiSMASH (Weber et al. 2015) identified 12 BGCs in the *Pseudomonas protegens pf-5* genome, including seven NRP and PK BGCs. SPAdes assembled each of them into a single contig with the exception of the pyoverdine NRP BGC (with eight A-domains), which was assembled into four contigs that revealed only seven A-domains (Figure 3, top left). In contrast, the domain restoration procedure in biosyntheticSPAdes succeeded in reconstructing two A-domains that were collapsed on a single edge by SPAdes (Figure 3, top right). The resulting scaffolding graph contains a single rural postman route that revealed the correct arrangement of A-domains (Figure 3, bottom). The reconstructed pyoverdine NRP BGC aligns to the *Pseudomonas protegens pf-5* genome with 99.9% identity.

Analyzing the CYANO dataset. Kleigrewe et al. 2015 assembled the CYANO dataset using SPAdes. metaSPAdes assembled the CYANO dataset into the assembly graph with 217,826 vertices and 116,066 edges (8454 of them are longer than 1 kb). metaSPAdes assembled the jamaicamide BGC with complexity 9 into a single contig but failed to assemble the hectochlorin BGC with complexity 5 into a single contig.

biosyntheticSPAdes extracted 781 BGC subgraphs, including 12 non-trivial BGC subgraphs with complexities 21, 20, 11, 9, 6, 6, 5, 5, 5, 5, 4, and 4. The hectochlorin BGC contains 22 domains (four A-domains, one AT-domain, four C-domains, one KS-domain, three KR domains and several others, one of them was also identified by HMMER as A-domain). biosyntheticSPAdes assembled the hectochlorin BGCs into a single contig (Figure 4) that aligns with the *Moorea producens JHB* genome with 99.9% identity. The jamaicamide BGC contains 42 domains (three A-domains, six AT-domains, four KR-domains, seven KS-domains, two C-domains, one TE-domain and several others). The jamaicamide scaffolding graph contains a single solid edge (usually, it means that the entire BGC was recovered after the repeat resolution step with exSPAnder).

Besides reconstructing the hectochlorin and the jamaicamides BGCs, biosyntheticSPAdes recovered sequences for 5 more putative NRP BGCs that were missed in previous studies (see Appendix: “Putative NRP BGCs in the CYANO dataset, Supplementary Figure S3 and Supplementary Figure S4).

Analyzing the MIBiG datasets. For each of 665 MIBiG datasets corresponding to a single known NRP or PK, we launched biosyntheticSPAdes on the SPAdes assembly graph. We also compared them with the

other popular assemblers: MEGAHIT v.1.1.3 (Li et al, 2015) and ABySS assembler v.2.1.0 (Simpson et al, 2009). For each assembler and each MiBiG dataset. the assembly was classified as successful if it meets the following criteria: (i) one of the contigs in the assembly covers more than 95% of the BGC and has at least 95% identity with the BGC being assembled and (ii) this contig has no misassemblies as identified by QUAST (Gurevich et al. 2013). biosyntheticSPAdes failed to successfully assemble only 11% of BGCs versus 22% for SPAdes, 35% for MEGAHIT and 34% for ABySS (Table 1). For 139 out of 665 BGCs with complexity >10, biosyntheticSPAdes failed to successfully assemble 22% of BGCs versus 58% for SPAdes, 79% for MEGAHIT and 83% for ABySS.

Analyzing the HMP datasets. To reconstruct BGCs in the human microbiome, we assembled each HMP dataset with biosyntheticSPAdes. We define the *biosynthetic capacity* of an assembly as the number of A and AT domains identified in this corresponding assembly. The biosynthetic capacity of the HMP datasets varies from 60 to over 400 across various human body sites (see Appendix: “Biosynthetic capacity of the HMP datasets” and Supplementary Table S4), suggesting that many HMP samples may encode over a dozen of NRP and PK BGCs. However, the amount of high-complexity BGC subgraphs suggests that sequencing depth in some datasets from the HMP project may be insufficient to capture the diversity of BGCs.

Below, we focus on analyzing the subpravingal plaque metagenome (dataset SRS013723) with large biosynthetic capacity. The assembly graph of this dataset contains 1540 BGC subgraphs, including seven non-trivial BGC subgraphs. We analyzed one of the complex BGC subgraphs with six predicted A-domains, five C-domains and two TE-domains that was assembled into six contigs. Figure 5 shows the BGC subgraph and two rural postman routes in the scaffolding graph generated by biosyntheticSPAdes. A nucleotide BLAST search of two predicted BGCs against the nt/rt database revealed only the short regions of similarity (less than 200 bp) with various *Pseudomonas* species, suggesting that Figure 5 represents a still unknown BGC. See Appendices “Biosynthetic capacity of the HMP datasets”, “Putative

NRP BGCs in subpravingal plaque datasets", Supplementary Figure S5 and Supplementary Table S5 for detailed analysis of the subpravingal plaque datasets.

DISCUSSION

While the human microbiome encodes natural products with great biomedical potential, little is known about these abundant small molecules, despite the fact that the human host is chronically exposed to them (Donia et al. 2014). One of the bottlenecks in discovering natural products from human and other metagenomes is deriving full-length BGCs from short metagenomics reads (Donia and Fischbach, 2015). This bottleneck negatively affects various genome mining efforts. Indeed, although the discovery of coelichelin (Challis et al. 2005) was one of the first successes of genome mining that was followed by the characterization of many NRPs from sequenced genomes, genome mining in fragmented assemblies remains challenging.

The discovery of the bioactive peptides teixobactin (Wilson et al. 2014) and polytheonamides (Freeman et al. 2012) marks a new era of natural product discovery from uncultivated bacteria. However, while various metagenomes serve as a rich source of natural products (Cragg et al. 2013, Katz et at. 2016), reconstructing complex BGCs from metagenomic assemblies is nearly impossible with short read sequencing technologies. Since gene prediction of BGC scattered between multiple contigs is challenging, the full-length BGC reconstruction is usually difficult without additional biological experiments and extensive manual analysis (Kleigrewe et al. 2015).

biosyntheticSPAdes is a step toward enabling high-throughput natural product discovery by coupling metagenomics and mass spectrometry projects using tools such as NRPquest (Mohimani et al. 2014). It represents the first automated pipeline for BGC reconstruction from genomic and metagenomic sequencing datasets that takes advantage of the assembly graph rather than individual contigs. While we demonstrated that biosyntheticSPAdes is able to recover long BGCs, it can also be extended to other

types of long and highly repetitive genes, such as 16S rRNA genes or insecticide toxins (Palma et al., 2014). Although biosyntheticSPAdes currently has the predefined options only for the most important classes of BGCs (NRPS, PKSs, and their fusions), we plan to create presets for other it can be extended for other BGCs with different domain compositions. A user can replace the default HMM-profiles with any profiles of interest, such as TPR-proteins, mucus-binding proteins, etc. However, we currently do not have plans to develop a version of SPAdes for generic operon prediction since it is not clear how to account for a wide diversity of genes within operons in general.

We emphasize that, similarly to all gene prediction tools, a putative BGC predicted by biosyntheticSPAdes may be incorrect and should be used with caution. In particular, the homology-based mode of biosyntheticSPAdes is most useful when one or more closely related reference genomes are available that have well-annotated BGCs. In the case when multiple feasible paths exist in the assembly graph, we recommend to experimentally verify biosyntheticSPAdes predictions, e.g., using targeted PCR amplification or matching against mass-spectrometry data. Also, peptidogenomics tools (Mohimani et al., 2014) can be applied to all feasible paths in the assembly graph rather than to a single highest-scoring path.

Third generation sequencing technologies have greatly improved isolate bacterial sequencing, thus turning BGC assembly into a relatively simple task. However, they have not yet had a large impact on metagenomic sequencing due to relatively high cost of long-read technologies and difficulties in assembly (no specialized long read metagenomic assembler has been released yet). Since most new natural products are analyzed through metagenomics (or mini-metagenomics) rather than isolate datasets, short reads remain the workhorse of genome mining for natural products.

Some researchers use hybrid approaches for metagenomics assemblies by combining short and long reads (Frank et al. 2016, Tsai et al. 2016). biosyntheticSPAdes is implemented in a manner that allows one to use new sequencing technologies as long as they are supported by the SPAdes pipeline. Since both SPAdes and metaSPAdes support hybrid datasets (Illumina + Pacific Bioscience/Oxford Nanopores), biosyntheticSPAdes can also assemble BGCs in hybrid datasets.

METHODS

Below we describe the six steps of the biosyntheticSPAdes pipeline (Figure 2) and illustrate them using reconstruction of the CALC BGC (Figure 1).

Step 1: Assembling genomic/metagenomic reads with SPAdes/metaSPAdes. BiosyntheticSPAdes starts with launching SPAdes (Bankevich et al. 2012) or metaSPAdes (Nurk et al. 2017) assemblers. SPAdes and metaSPAdes first construct a *de Bruijn graph* (Compeau et al. 2011) of all reads and subsequently perform various graph simplification procedures (e.g., *bubble collapsing* and *tip removal*) to transform it into an *assembly graph*. Both SPAdes and metaSPAdes use exSPAnder (Prjibelsky et al. 2014) to utilize the read-pair information for repeat resolution and scaffolding in the assembly graph.

Step 2: Identifying domain-edges in the assembly graph. The first step towards reconstructing the nucleotide sequence of a BGC is reconstruction of the arrangement of its biosynthetic domains. In many cases, this arrangement alone provides sufficient information for predicting the structure of the core scaffold of a natural product encoded by the BGC.

To identify edges harboring biosynthetic domains in the assembly graph, contigs generated by SPAdes/metaSPAdes are searched for the domain motifs using HMMER (Zhang et al. 2003, Eddy, 2011).

For illustration purposes, here we analyze only A-domains. After mapping contigs back to the assembly graph, biosyntheticSPAdes identifies the positions of all detected domains in the assembly graph (Figure 1, top). Mapping the A-domains from the CALC BGC back to the assembly graph reveals that three A-domains (4, 5, and 7) map to the same positions on a single edge of the assembly graph. The edge harboring these positions has approximately three times higher coverage than the average coverage of edges that contain only a single copy of an A-domain. Supplementary Figure S1 illustrates that these three domains are similar to each other, and share identical repeats of length ≈100 bp and longer. Sequences of these domains are collapsed during assembly, because the assembly graph was constructed from *k*-mers that are shorter than 100 nucleotides.

Step 3: Extracting BGC subgraphs from the assembly graph. BGCs contain various domains and multiple biosynthetic genes in close proximity to each other. Analysis of all complete NRP BGCs from the MIBiG repository of BGCs (Medema et al. 2015) revealed that the distances between consecutive NRPS- or PKS-related domains do not exceed 20 kb in 95% of the cases (Supplementary Figure S2).

Hence, we consider all edges in the assembly graph within 10 kb from the positions of domains on the domain edges identified in the previous step to capture all consecutive domains separated by at most 20 kb. The subgraph of the assembly graph formed by these edges, referred to as the *BGC assembly graph*, usually consists of multiple connected components, where each component, referred to as a *BGC subgraph*, usually corresponds to a single BGC. For example, four NRP BGCs in *S. coelicolor* genome are represented by four different connected components of the BGC assembly graph. However, in some cases a single component of the BGC assembly graph may combine multiple BGCs, particularly when these BGCs share very similar domains with identical sequences exceeding the maximum default *k*-mers size in SPAdes. The *complexity of the BGC subgraph* is defined as the total number of A-domains and AT-domains in this subgraph. We define *non-trivial BGC subgraphs* as BGC subgraphs of complexity at least 3.

The BGC assembly graph for *S.coelicolor* consists of 24 BGC subgraphs. Three of them are non-trivial BGC subgraphs with complexities 9 (for the CALC BGC), 4, and 3. The BGC subgraph corresponding to the CALC BGC with 11 A-domains revealed only 9 A-domains, since three A-domains were collapsed into a single edge.

Step 4: Restoring collapsed domains in the assembly graph. Figure 1 reveals a limitation of existing assemblers (*repeat collapsing*) that negatively affects gene prediction tools: three A-domains sharing long identical segments are collapsed into a single edge in the assembly graph. As a result, valuable information about the differences between these A-domains is lost (Supplementary Figure S1). This effect is amplified in metagenomics assemblies since they aggressively collapse bubbles to improve contiguity of the assembly (Nurk et al., 2017), particularly in the case of metagenomes containing similar strains. A side effect of the bubble collapsing procedure is collapsing similar domains, which leads to a high number of mismatches and indels in reconstructed BGC sequences (referred to as an "assembly deterioration").

This limitation of the existing assemblers can be remedied by restoring subtle variations in the collapsed repeats to enable better repeat resolution. Since SPAdes and metaSPAdes provide map each read to the assembly graph, we consider all reads mapped to edges of all BGC subgraphs and compute the median depth of coverage of each edge. Given an edge with coverage *cov* in a BGC subgraph, we extract all *k*-mers from the reads mapped to this edge. A *k*-mer is defined as *solid* if it does not belong to the edge but appears in at least α**cov* reads mapped to this edge (the default value α=0.2). Solid *k*-mers reveal variations in repeats (rather than sequencing errors), as the expected frequency of erroneous *k*-mers is typically below α**cov*. We define a path formed by solid *k*-mers as a *solid bubble* if it forms an alternative path in a BGC subgraph. We restore all such solid bubbles in a BGC subgraph and rerun the exSPAnder repeat resolution on the modified BGC subgraphs with restored solid bubbles. We emphasize that we applied the domain restoration step to the domain edges in the BGC subgraphs only since

applying it to the entire assembly graph leads to deterioration of the assembly and reduced N50 statistics.

Note that the consensus sequence of the edge harboring three similar but not identical A-domains in the CALC assembly (Figure 6) differs from the sequences of each of these A-domains. Therefore, it provides slightly inaccurate sequences for each of these three domains. However, after the domain restoration procedure, these three A-domains correspond to three different and 100% accurate consensus sequences. In some cases, the domain restoration procedure even enables exSPAnder to utilize the restored variations between domains for further repeat resolution by utilizing variations between long imperfect repeats. We note that although the described bubble restoration procedure has a potential to resolve close strains in metagenomics assemblies, it has not been implemented in metaSPAdes yet.

Running exSPAnder on the modified BGC subgraph with restored bubbles often results in a more accurate estimate of the total number of domains (Figure 6). In contrast to the initial BGC subgraph with only 9 identified A-domains for the CALC BGC, all 11-A-domains are now captured in 5 resulting contigs in the modified BGC subgraph.

Step 5: Constructing the scaffolding graph. We represent each domain-containing contig as an isolated *solid edge* in the *scaffolding graph* (Figure 7). Given solid edges *e* and *e'*, we connect the ending vertex of *e* with the starting vertex of *e'* by a *dashed edge* if the last domain on *e* and the first domain of *e'* are close in the BGC assembly graph, i.e., the distance between them is below 10 kb. Given a directed graph with solid and dashed edges, the *Rural Postman Problem* is to find a rural postman route, i.e., a path visiting all solid edges of the graph (Orloff, 1974).

Step 6: Constructing putative BGCs by solving the Rural Postman Problem. Inferring the

arrangement of domains in an NRP BGC is crucial for identifying the NRP encoded by this NRPS. Since each NRP synthetase corresponds to a rural postman routes in the scaffolding graph, biosyntheticSPAdes searches for all rural postman routes in the scaffolding graph using a brute force algorithm (most scaffolding subgraphs have less than 20 vertices). Figure 7 shows two rural postman routes in the CALC scaffolding graph.

Some bacterial genomes contain 100% identical domains that are collapsed into a single edge even after domain restoration. As the result, a rural postman route may visit the collapsed solid edges in some scaffolding graphs multiple times. For each solid edge in the scaffolding graph, the approximate number of times it should be traversed is defined by the ratio of the coverage of the corresponding domain-edge in the BGC subgraph to the median coverage across all edges of the BGC subgraph.

As Figure 7 illustrates, biosyntheticSPAdes may output multiple arrangements of A-domains, each arrangement corresponding to a rural postman route. For each rural postman route, biosyntheticSPAdes reconstructs a path in the BGC assembly graph corresponding to this route and its nucleotide sequence. Dashed edges in a rural postman route may correspond to multiple paths in the BGC assembly graph, and we report the path with the length closest to any of distances from the set of 550, 1500 and 2400 bp, the values of the three pronounced peaks in the distribution of the distances between consecutive domains in known NRPSs (Supplementary Figure S2).

biosyntheticSPAdes and NRPquest for PNP reconstruction. Even when biosyntheticSPAdes fails to assemble a BGC into a single contig, it typically reduces the number of contigs as compared to SPAdes, e.g., outputs two contigs A and B without providing one of two possible orders to concatenate these contigs (B after A or A after B). This feature is important for natural product researchers since they often perform additional experiments to reconstruct the correct order of contigs (Kleigreve, 2015). For example, in the case of NRP BGC, one can generate all possible concatenates, predict putative NRPs for each

concatenate, and match a spectral dataset against all putative NRPs to find a concatenate with the best match. Appendix: “Output format of biosyntheticSPAdes” specifies the details of the biosyntheticSPAdes output. Appendix: “Coupling biosyntheticSPAdes and NRPquest for PNP reconstruction” presents an example of combining genomic and mass spectrometry data to infer the correct arrangement of A-domains.

Extending biosyntheticSPAdes from NRP BGCs to other BGCs. In addition to the A-domains, biosyntheticSPAdes analyzes other domains in NRP BGCs such as *C–condensation* domains (*C-domains*) and *thioesterase* domains (*TE-domains*), among others. Moreover, biosyntheticSPAdes is not limited to NRP BGCs and also works with BGCs encoding PKS BGCs (Robinson, 1991). PKSs are built from various domains including *acyltransferase* domains (AT-domains), *keto-synthase* domains (KS-domains), *keto-reductase* domains (KR-domains) and *acyl carrier protein domains (ACP-domains).*

Reference-based BGC ranking algorithm. When a database of reference genomes is available, it can help to predict the correct order of contigs by identifying a genome with a similar BGC. This is especially relevant when assembling genomes that are related to an already sequenced species, or during studies of microbial communities from which individual strains have been isolated and sequenced. biosyntheticSPAdes includes a pipeline that matches all possible orders of multiple putative BGC sequences to gene clusters in antiSMASH-DB (Blin et al, 2016) and ranks them based on how well the order of the matching domains corresponds to the domain order in the most similar BGC.

Note that the reference-based BGC ranking algorithm is an optional module in biosyntheticSPAdes that should be called only in cases when there is more than one plausible path in the assembly graph. In most of our test cases, biosyntheticSPAdes leads to a single plausible path through the assembly graph, and thus a single BGC architecture. In all such cases, reference genomes are not required to infer the correct assembly.

In the case when a BGC-subgraph is not resolved into a single BGC, biosyntheticSPAdes generates multiple putative BGCs (*pBGCs*) and ranks them based on their similarity to BGCs from reference genomes from antiSMASH-DB (Blin et al. 2016). Each pBGC is compared to each reference BGCs (*rBGCs*) and scored according to the similarity between the pBGC and the rBGCs with respect to sequence similarity, domain composition, and domain order. See Appendices: "Reference-based putative BGC ranking algorithm", "Ranking putative BGCs from *Streptomyces coelicolor* A3(2) and *Streptomyces avermitilis* MA-4680", Supplementary Figures S6, S7, S8, and Supplementary Tables S6, S7 for details.

Software Availability

biosyntheticSPAdes will be included in the next version of the SPAdes toolkit available from http://cab.spbu.ru/software/spades starting from version 3.14. The pre-release version, that was used for benchmarking in this paper and the biosyntheticSPAdes ranking pipeline, is available in Supplemental material. BiosyntheticSPAdes source code is alternatively available from http://dx.doi.org/10.6084/m9.figshare.6948260.v2 and the biosyntheticSPAdes ranking pipeline is alternatively available from https://git.wur.nl/medema-group/biosyntheticSpadesRankingPipeline.

Acknowledgements

We are grateful to Alexey Gurevich, Sergey Nurk, Bahar Behsaz, and Jeremy Owen for useful discussions and assistance with data analysis. A.K. was supported by the Russian Science Foundation (grant 19-14-00172). V.T. is supported by the research program NWO-Groen, which is jointly funded by the Netherlands Organization for Scientific Research (NWO), BASF SE and Baseclear BV (project ALWGR.2015.1). M.H.M. is supported by VENI grant 863.15.002 from The Netherlands Organization for Scientific Research (NWO). DM was supported by the Tri-Institutional Training Program in Computational Biology and Medicine (NIH grant 1T32GM083937).

Disclosure Declaration

Authors have no conflicts to report.

REFERENCES

Bentley, S. D., Chater, K. F., Cerdeno-Tarraga, A. M., Challis, G. L., Thomson, N. R., James, K. D., & Bateman, A. (2002). Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3

(2). *Nature*, *417*(6885), 141-147.

Besemer, J., & Borodovsky, M. (2005). GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Research, 33* (suppl. 2), W451-W454.

Blin, K., Medema, M. H., Kottmann, R., Lee, S. Y., & Weber, T. (2016). The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. *Nucleic Acids Research*, 45, D555-D559.

Blin, K., Medema, M. H., Kazempour, D., Fischbach, M. A., Breitling, R., Takano, E., & Weber, T. (2013). antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Research*, *41*(W1), W204-W212.

Cane, D. E., & Walsh, C. T. (1999). The parallel and convergent universes of polyketide synthases and nonribosomal peptide synthetases. *Chemistry & Biology*, *6*(12), R319-R325.

Challis, G. L., & Ravel, J. (2000). Coelichelin, a new peptide siderophore encoded by the Streptomyces coelicolor genome: structure prediction from the sequence of its non-ribosomal peptide synthetase. *FEMS Microbiology Letters*, *187*(2), 111-114.

Chu, J., Vila-Farres, X., Inoyama, D., Ternei, M., Cohen, L. J., Gordon, E. A., ... & Jaskowski, M. (2016). Discovery of MRSA active antibiotics using primary sequence from the human microbiome. *Nature Chemical Biology*, *12*(12), 1004.

Coates RC, Podell S, Korobeynikov A, Lapidus A, Pevzner P, Sherman DH, Allen EE, Gerwick L., Gerwick WH. (2014) Characterization of cyanobacterial hydrocarbon composition and distribution of biosynthetic pathways. *PLoS One*. 9(1):e85140.

Compeau, P. E., Pevzner, P. A., & Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, *29*(11), 987-991.

Cragg, G. M., & Newman, D. J. (2013). Natural products: a continuing source of novel drug leads. *Biochimica et Biophysica Acta*, *1830*(6), 3670-3695.

Delcher, A. L., Bratke, K. A., Powers, E. C., & Salzberg, S. L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, *23*(6), 673-679.

Donia, M. S., Cimermancic, P., Schulze, C. J., Brown, L. C. W., Martin, J., Mitreva, M., Clardy J., Linington R.G., and Fischbach, M. A. (2014). A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. *Cell*, *158*(6), 1402-1414

Donia, M. S., and Fischbach, M. A. (2015). Small molecules from the human microbiota. *Science*, 349:1254766.

Eddy, S. R. Accelerated profile HMM searches. *PLoS computational biology* 7.10 (2011): e1002195.

Edwards, D. J., Marquez, B. L., Nogle, L. M., McPhail, K., Goeger, D. E., Roberts, M. A., & Gerwick, W. H. (2004). Structure and biosynthesis of the jamaicamides, new mixed polyketide-peptide neurotoxins from the marine cyanobacterium Lyngbya majuscula. *Chemistry & Biology*, *11*(6), 817-833.

Frank, J.A., Pan, Y., Tooming-Klunderud, A., Eijsink, V.G., McHardy, A.C., Nederbragt, A.J. and Pope, P.B., 2016. Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. *Scientific reports*, *6*, 25373.

Freeman, M. F., Gurgui, C., Helf, M. J., Morinaka, B. I., Uria, A. R., Oldham, N. J., ... & Piel, J. (2012). Metagenome mining reveals polytheonamides as posttranslationally modified ribosomal peptides. *Science*, *338*(6105), 387-390.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072-1075.

Gurevich, A., Mikheenko, A., Shlemov, A., Korobeynikov, A., Mohimani, H., & Pevzner, P. A. (2018). Increased diversity of peptidic natural products revealed by modification-tolerant database search of mass spectra. *Nature microbiology*, *3*(3), 319.

Hadjithomas, M., Chen, I. M. A., Chu, K., Ratner, A., Palaniappan, K., Szeto, E., ... & Ivanova, N. N. (2015). IMG-ABC: a knowledge base to fuel discovery of biosynthetic gene clusters and novel secondary metabolites. *MBio*, 6(4), e00932-15.

Hobbs, G. A. I. C., Obanye, A. I., Petty, J., Mason, J. C., Barratt, E., Gardner, D. C., ... & Oliver, S. G. (1992). An integrated approach to studying regulation of production of the antibiotic methylenomycin by *Streptomyces coelicolor* A3 (2). *Journal of Bacteriology*, *174*(5), 1487-1494.

Huang, W., Li, L., Myers, J. R., & Marth, G. T. (2012). ART: a next-generation sequencing read simulator. *Bioinformatics*, *28*(4), 593-594.

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*(1), 119.

Ikegami, T., Inatsugi, T., Kojima, I., Umemura, M., Hagiwara, H., Machida, M., Asai, K. (2015 Hybrid *De Novo* Genome Assembly Using MiSeq and SOLiD Short Read Data *PLoS ONE* 10(4): e0126289.

Katz, M., Hover, B. M., & Brady, S. F. (2016). Culture-independent discovery of natural products from soil metagenomes. *Journal of Industrial Microbiology & Biotechnology*, *43*(2-3), 129-141.

Kersten, R. D., Yang, Y.-L., Xu, Y., Cimermancic, P. and Nam, S.J., Fenical, W., Fischbach, M.A., Moore, B.S. Dorrestein, P.C. (2011) A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nat. Chem. Biol.* 7, 794–802

Klassen, J. L., & Currie, C. R. (2012). Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation. *BMC genomics*, *13*(1), 14.

Kleigrewe, K., Almaliti, J., Tian, I. Y., Kinnel, R. B., Korobeynikov, A., Monroe, E. A., ... & Gerwick, L. (2015). Combining mass spectrometric metabolic profiling with genomic analysis: A powerful approach for discovering natural products from cyanobacteria. *Journal of Natural Products*, *78*(7), 1671-1682.

Li, D., Liu, C. M., Luo, R., Sadakane, K., & Lam, T. W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, *31*(10), 1674-1676.

Liu, C. M., McDaniel, L. E., & Schaffner, C. P. (1972). Fungimicyn biogenesis of its aromatic moiety. *The Journal of Antibiotics*, *25*(3), 187-188.

Magnolo, S. K., Leenutaphong, D. L., DeModena, J. A., Curtis, J. E., Bailey, J. E ., Galazzo, J. L., & Hughes, D. E. (1991). Actinorhodin production by *Streptomyces coelicolor* and growth of *Streptomyces lividans* are improved by the expression of a bacterial hemoglobin. *Biotechnology, 9*(5), 473-476.

Marquez, B. L., Watts, K. S., Yokochi, A., Roberts, M. A., Verdier-Pinard, P., Jimenez, J. I., ... & Gerwick, W. H. (2002). Structure and absolute stereochemistry of hectochlorin, a potent stimulator of actin assembly. *Journal of Natural Products*, *65*(6), 866-871.

Medema, M. H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M. A., ... & Breitling, R. (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research*, *39* (suppl. 2), W339-W346.

Medema, M.H., Cimermancic P., Sali A., Takano E., Fischbach, M.A. (2014) A Systematic Computational Analysis of Biosynthetic Gene Cluster Evolution: Lessons for Engineering Biosynthesis. *PLOS Computational Biology 10*, e1004016

Medema, M. H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J. B., Blin, K., ... & Cruz-Morales, P. (2015). Minimum information about a biosynthetic gene cluster. *Nature Chemical Biology*, *11*(9), 625-631.

Methé, B. A., Nelson, K. E., Pop, M., Creasy, H. H., Giglio, M. G., Huttenhower, C., ... & Chinwalla, A. T. (2012). A framework for human microbiome research. *Nature*, *486*(7402), 215.

Medema MH, Paalvast Y, Nguyen DD, Melnik A, Dorrestein PC, Takano E, et al. (2014) Pep2Path: automated mass spectrometry-guided genome mining of peptidic natural products. *PLoS Comput Biol* 10: e1003822.

Mohimani, H., Liu, W. T., Kersten, R. D., Moore, B. S., Dorrestein, P. C., & Pevzner, P. A. (2014). NRPquest: coupling mass spectrometry and genome mining for nonribosomal peptide discovery. *Journal of Natural Products*, *77*(8), 1902-1909.

Mohimani, H., Kersten, R. D., Liu, W. T., Wang, M., Purvine, S. O., Wu, S., ... & Pevzner, P. A. (2014). Automated genome mining of ribosomal peptide natural products. *ACS Chemical Biology*, *9*(7), 1545-1551.

Mohimani, H., Gurevich, A., Mikheenko, A., Garg, N., Nothias, L. F., Ninomiya, A., ... & Pevzner, P. A. (2017). Dereplication of peptidic natural products through database search of mass spectra. *Nature Chemical Biology*, *13*(1), 30-37.

Mukherjee, S., Seshadri, R., Varghese, N. J., Eloe-Fadrosh, E. A., Meier-Kolthoff, J. P., Göker, M., ... & Yoshikuni, Y. (2017). 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nature Biotechnology*. 35(7):676-683.

Newman, D.J., Cragg, G.M. Natural Products as Sources of New Drugs from 1981 to 2014 (2016) *J. Natural Products*, *79*, 629–661

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, *27*(5), 824-834.

Orloff, C. S. (1974). A fundamental problem in vehicle routing. *Networks*, *4*(1), 35-64.

L. Palma, D. Muñoz, C. Berry, J. Murillo, P. Caballero (2014) *Bacillus thuringiensis* toxins: an overview of their biocidal activity. *Toxins*, 6, 3296-3325.

Pati, A., Ivanova, N. N., Mikhailova, N., Ovchinnikova, G., Hooper, S. D., Lykidis, A., & Kyrpides, N. C. (2010). GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. *Nature Methods*, *7*(6), 455-457.

Paulsen, I. T., Press, C. M., Ravel, J., Kobayashi, D. Y., Myers, G. S., Mavrodi, D. V., ... & Dodson, R. J. (2005). Complete genome sequence of the plant commensal *Pseudomonas fluorescens* Pf-5. *Nature Biotechnology*, *23*(7), 873-878.

Robinson, J. A. (1991). Polyketide synthase complexes: their structure and function in antibiotic biosynthesis. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *332*(1263), 107-114.

Röttig, M., Medema, M. H., Blin, K., Weber, T., Rausch, C., & Kohlbacher, O. (2011). NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Research*, *39 (suppl.2)*, W362-W367.

Prjibelski, A. D., Vasilinetc, I., Bankevich, A., Gurevich, A., Krivosheeva, T., Nurk, S., ... & Pevzner, P. A. (2014). ExSPAnder: a universal repeat resolver for DNA fragment assembly. *Bioinformatics*, *30*(12), i293-i301.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., & Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Research*, *19*(6), 1117-1123.

Stachelhaus, T., Mootz, H. D., & Marahiel, M. A. (1999). The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chemistry & Biology*, *6*(8), 493-505.

Strieker, M., Tanović, A., & Marahiel, M. A. (2010). Nonribosomal peptide synthetases: structures and dynamics. *Current opinion in structural biology*, *20*(2), 234-240.

Takano, E., Gramajo, H. C., Strauch, E., Andres, N., White, J., & Bibb, M. J. (1992). Transcriptional regulation of the redD transcriptional activator gene accounts for growth-phase-dependent production of the antibiotic undecylprodigiosin in *Streptomyces coelicolor* A3 (2). *Molecular Microbiology*, *6*(19), 2797-2804.

Tsai, Y.C., Conlan, S., Deming, C., Segre, J.A., Kong, H.H., Korlach, J., Oh, J. and NISC Comparative Sequencing Program, 2016. Resolving the complexity of human skin metagenomes using single-molecule sequencing. *MBio*, *7*(1),.e01948-15.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C., Knight, R., & Gordon, J. I. (2007). The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature*, *449*(7164), 804.

Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Bruccoleri, R., ... & Breitling, R. (2015).

antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Research*, *43*(W1), W237-W243.

Wilson, M. C., Mori, T., Rückert, C., Uria, A. R., Helf, M. J., Takada, K., ... & Rinke, C. (2014). An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature*, *506*(7486), 58-62.

Zhang, Z., & Wood, W. I. (2003). A profile hidden Markov model for signal peptides generated by HMMER. *Bioinformatics*, *19*(2), 307-308.

FIGURE LEGENDS

Figure 1. Subgraph of the assembly graph of *S. coelicolor* corresponding to the CALC NRP BGC. (Top) Edges of the assembly graph traversed by the CALC BGC. Nodes of the assembly graph are shown as white circles. After applying exSPAnder, the CALC BGC remains scattered over ten scaffolds. Three of them are shown as red, blue, and green paths through the assembly graph, the remaining seven consist of a single edge each (shown in black and marked with letters a through g). The positions of eleven A-domains (with their indices) along the CALC BGC are shown by violet boxes. Edges with low and high coverage by reads are shown as thin and thick edges, respectively. The edge harboring three A-domains 4, 5, and 7 has approximately triple coverage by reads as compared to other domain-harboring edges. The 11 A-domains in CALC are split over three NRP synthetases with 6, 3, and 2 A-domains, respectively. (Middle) A simplified representation of the graph with all short edges (shorter than 300 bp) contracted into single vertices. The two contracted subgraphs of the assembly graph (formed by short edges) are represented by yellow vertices. The brown dashed path illustrates how the CALC NRP synthetase traverses the contracted assembly graph. (Bottom) The bubble restoration procedure described below transforms the collapsed edge harboring three A-domains (A-domains 4, 5, and 7) into three edges, each of them harboring a single A-domain. Applying exSPAnder to the modified assembly graph results in seven scaffolds that differ from scaffolds before bubble restoration (shown as red, blue, green, and orange paths as well as three black edges). Grey squares show the starting and ending positions of the CALC BGC.

Figure 2. The biosyntheticSPAdes pipeline. Six steps of the biosyntheticSPAdes pipeline: (i) assembling genomic/metagenomic reads with SPAdes/metaSPAdes, (ii) searching for edges harboring biosynthetic domains in the assembly graph, (iii) extracting biosynthetic gene cluster subgraphs from the assembly graph, (iv) restoring the collapsed domains in the BGC-subgraphs, (v) constructing the scaffolding graph, and (vi) generating putative BGC by solving the Rural Postman Problem in the scaffolding graph.

Figure 3. Subgraph of the assembly graph of *Pseudomonas protegens* Pf-5 corresponding to the pyoverdine NRP BGC. (Top left) The pyoverdine BGC is scattered over four scaffolds in the SPAdes assembly. Two scaffolds traversing single edges are shown by black color and two scaffolds traversing multiple edges are shown by red and green colors. The repeat edges traversed by both red and green scaffolds are shown by brown color. Edges with low and high depth of coverage by reads are shown as thin and thick edges, respectively. Some A-domains span multiple edges (starting and ending positions of such domains are shown with dashed lines). (Top right) The domain restoration procedure restored two A-domains (5 and 6) in the assembly (SPAdes collapsed these domains into a single edge). Four scaffolds in the assembly graph are shown by red, green, blue and black colors. (Bottom) The scaffolding graph of the pyoverdine BGC with a single rural postman route (dashed edges in this route are shown in blue).

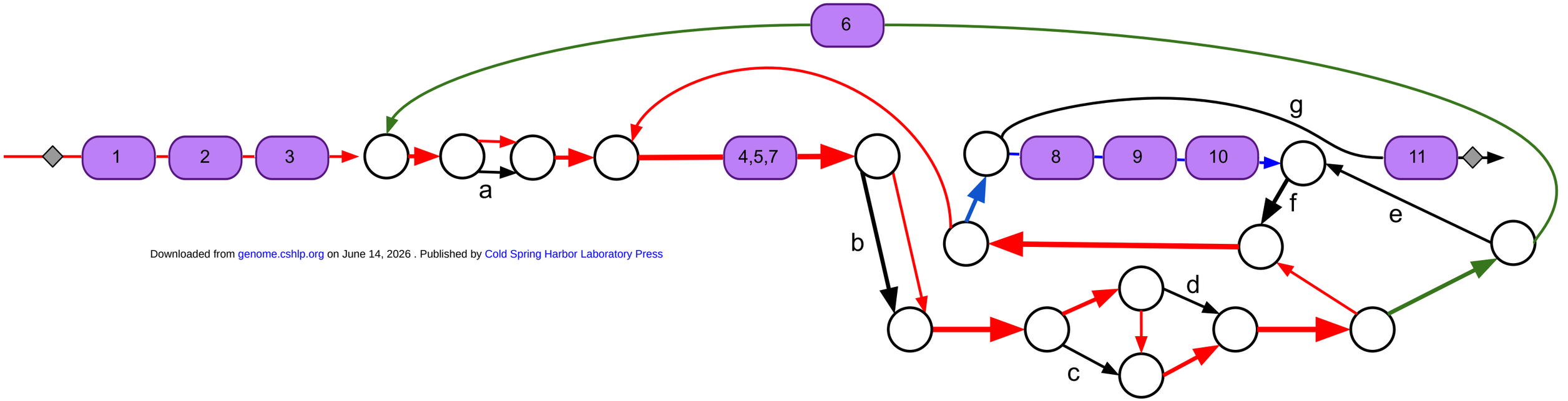
Figure 4. biosyntheticSPAdes assembly of the hectochlorin BGCs (the CYANO dataset). (Top) The subgraph of the assembly graph corresponding to the hectochlorin BGC. metaSPAdes assembly results in 4 scaffolds shown by a red path, a green path, and two black edges. The repeat edges traversed by both red and green scaffolds are shown by the brown color. The domain restoration procedure had no effect on this graph. (Bottom) The scaffolding graph of the hectochlorin BGC has only one rural postman route that revealed the correct domain order.

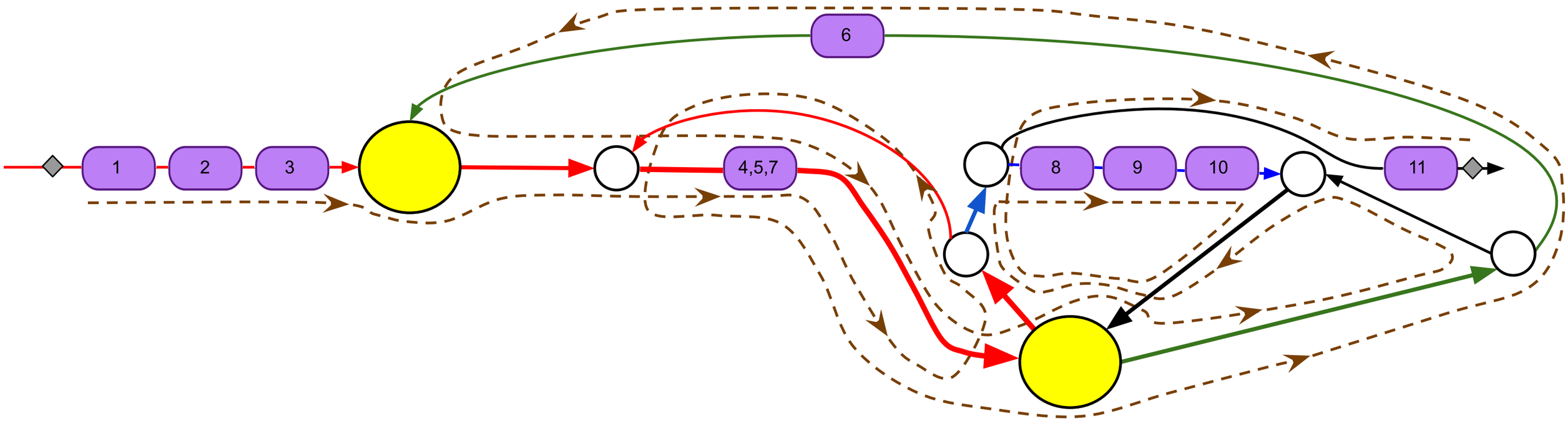
Figure 5. The BGC subgraph and the scaffolding graph for the subpravingal plaque metagenome (SRS013723) in the HMP dataset. (1,2) The BGC subgraph and the scaffolding graph, (3,4) Two rural postman routes in the scaffolding graph. The duplicated C-domain is highlighted with red border and is traversed twice in the rural postman routes. The numbers labeling the dashed edges indicate their order in the resulting tour. (5) Since biosyntheticSPAdes and antiSMASH use different thresholds and filtering options antiSMASH identified only 5 (rather than 6) A-domains in the NRP BGC predicted by biosyntheticSPAdes. Three most likely amino acids for each A-domain are shown along with their NRPSpredictor2 (Röttig et al. 2011) scores for the first of two rural postman routes.

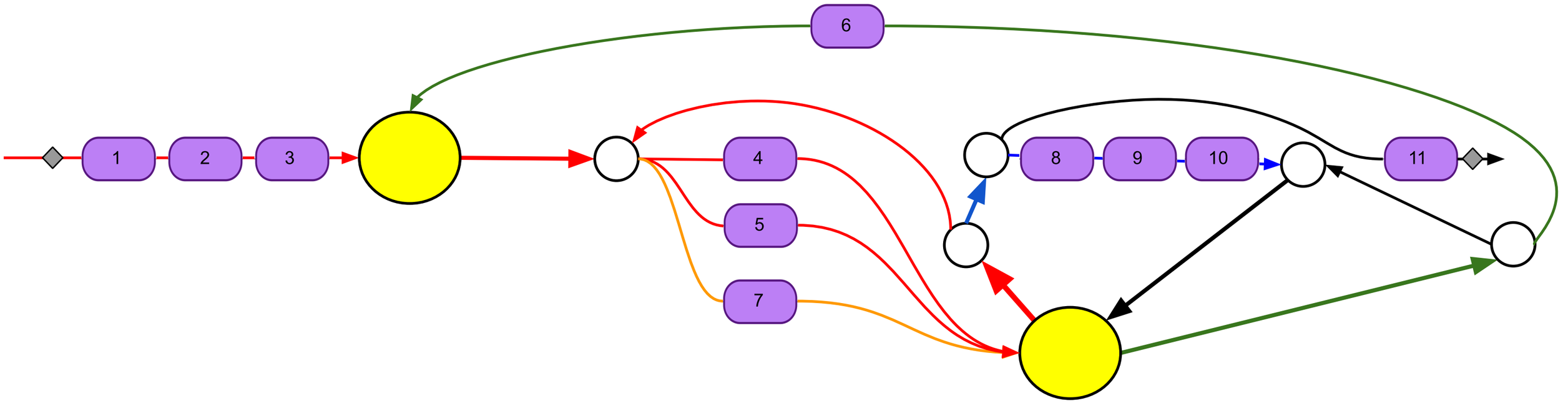
Figure 6. Effect of bubble restoration on the reconstruction of the CALC BGC. Schematic representation of repeat collapsing and consensus deterioration in the case of the CALC BGC assembly. While SPAdes outputs a single (and incorrect) consensus sequences of all three collapsed A-domains, these three sequences are not identical. In contrast, biosyntheticSPAdes utilized restored domains and reconstructed their distinct sequences with 100% accuracy (as compared to 99.6% accuracy for SPAdes). Numbers near dashed vertical lines represent the column numbers in the multiple alignment of three A-domain.

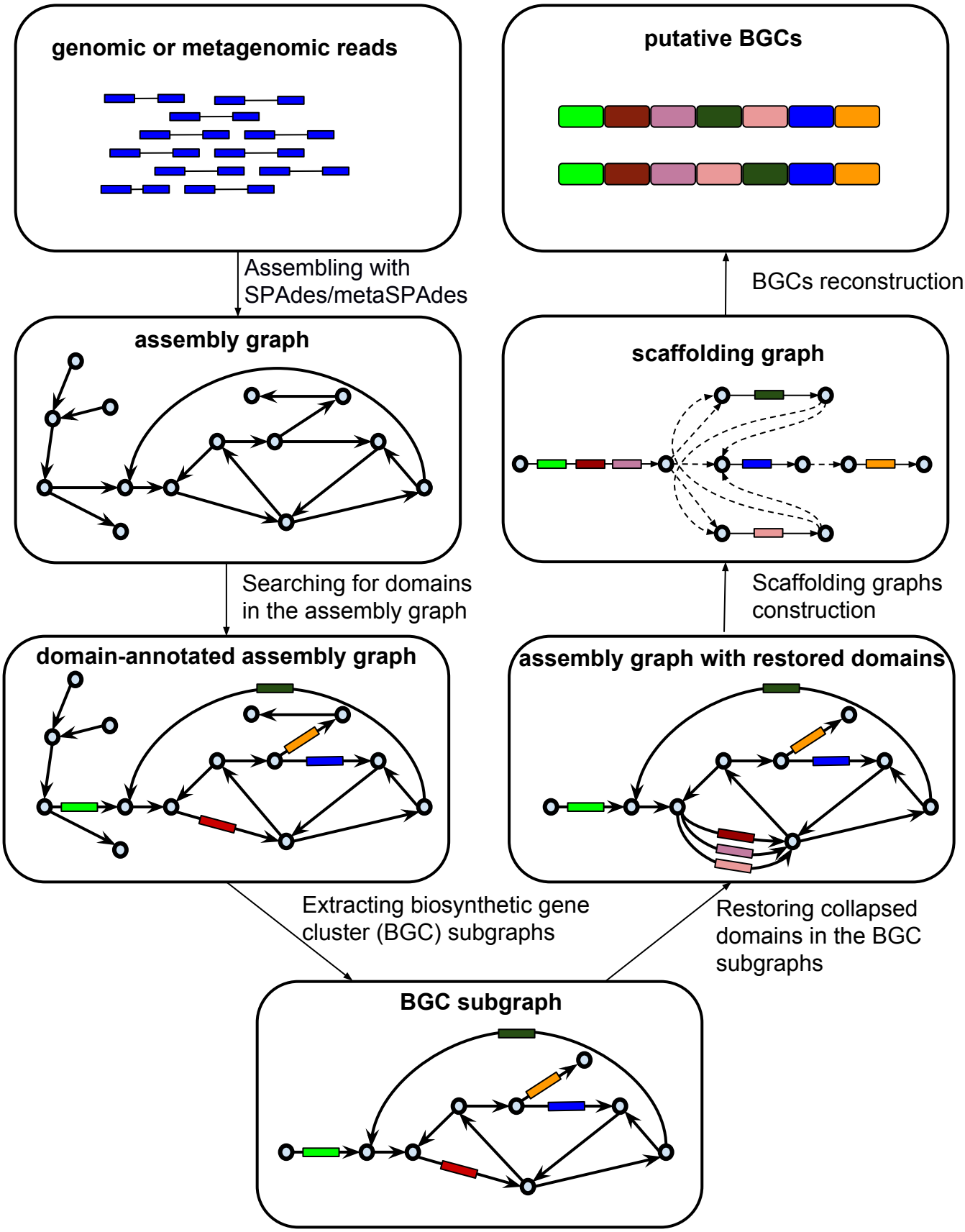
Figure 7. The scaffolding graph of the CALC BGC. (Left) Five solid edges in the scaffolding graph correspond to 5 contigs shown in Figure 4 (bottom) that contain A-domains. These contigs are shown as a red edge (A-domains 1, 2, 3, 4, and 5), a green edge (A-domain 6), a pink edge (A-domain 7), a blue edge (A-domains 8, 9, and 10), and a black edge (A-domain 11). Eight dashed edges in the scaffolding graph connect solid edges that contain closely located domains in the BGC subgraph. (Right) Two rural postman routes in the CALC scaffolding graph. First tour contains all violet dashed edges and results in the (1, 2, 3, 4, 5, **6, 7**, 8, 9, 10, 11) arrangement of A-domains while the second tour contains all brown dashed edges and results in the (1, 2, 3, 4, 5, **7, 6**, 8, 9, 10, 11) arrangement of A-domains.

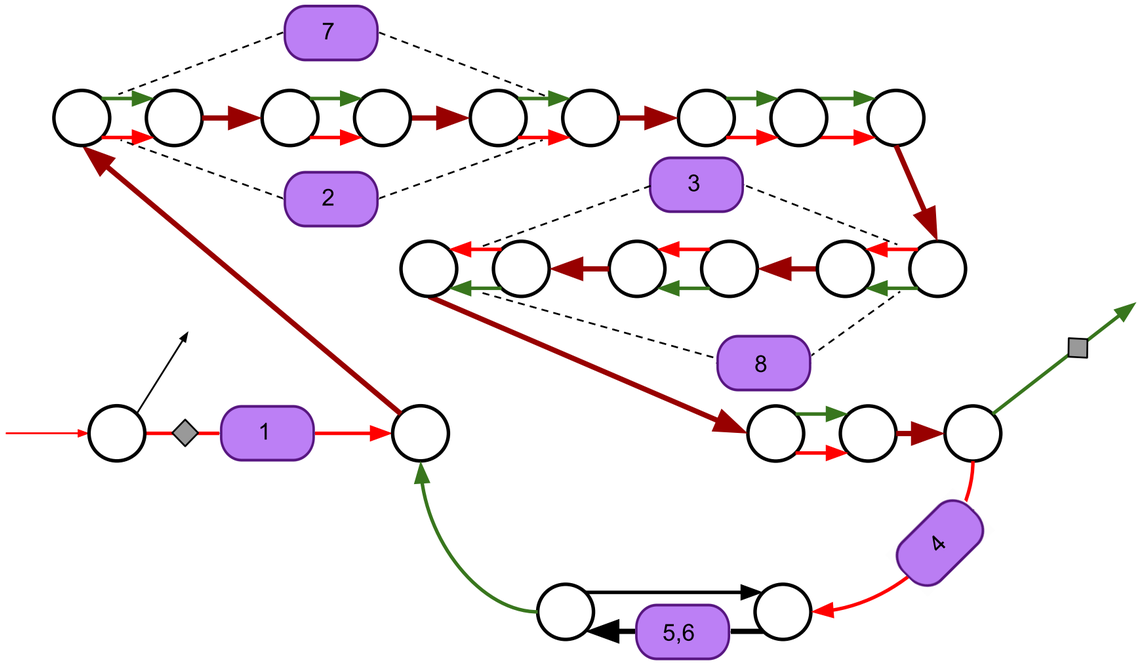
Table 1. Results of SPAdes, biosyntheticSPAdes, MEGAHIT and ABySS assemblies on 665 MiBIG datasets.

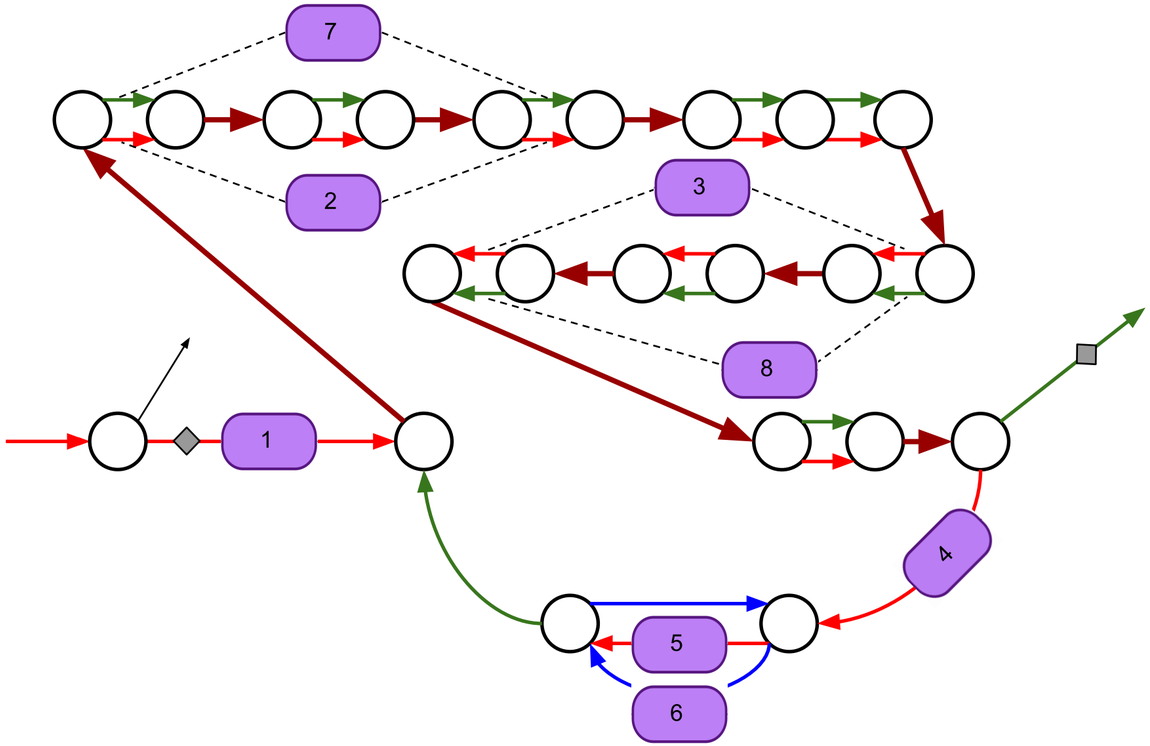


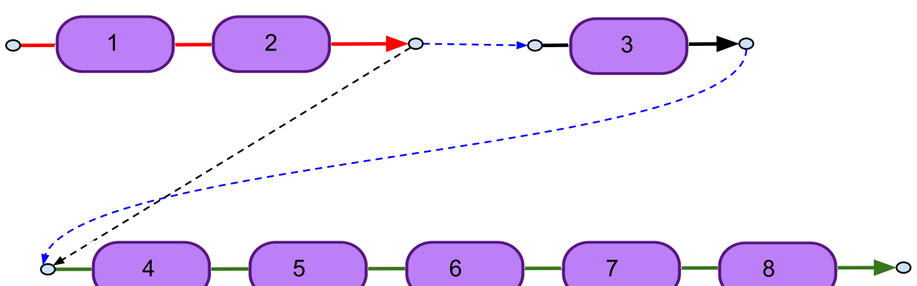


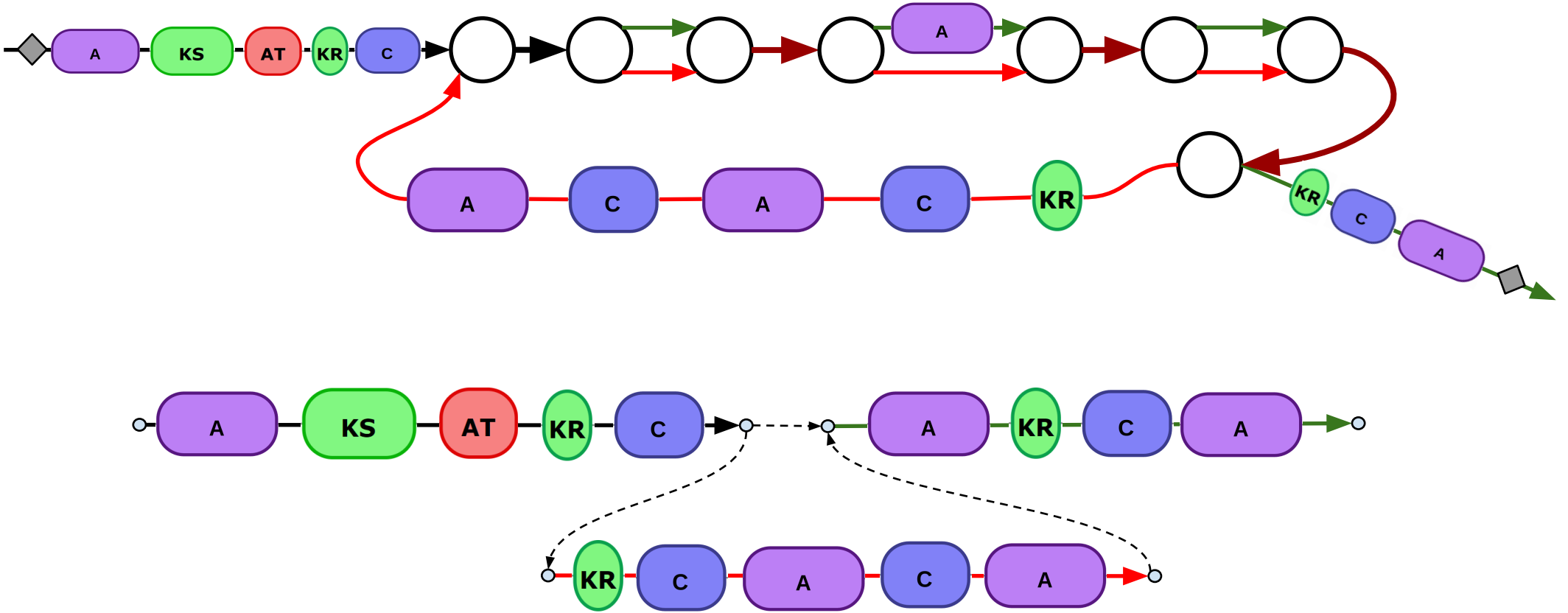



genomic or metagenomic reads
putative BGCs
Assembling with
SPAdes/metaSPAdes
BGCs reconstruction
assembly graph
scaffolding graph
Searching for domains
in the assembly graph
Scaffolding graphs
construction
domain-annotated assembly graph
assembly graph with restored domains
Extracting biosynthetic gene
cluster (BGC) subgraphs
Restoring collapsed
domains in the BGC
subgraphs
BGC subgraph

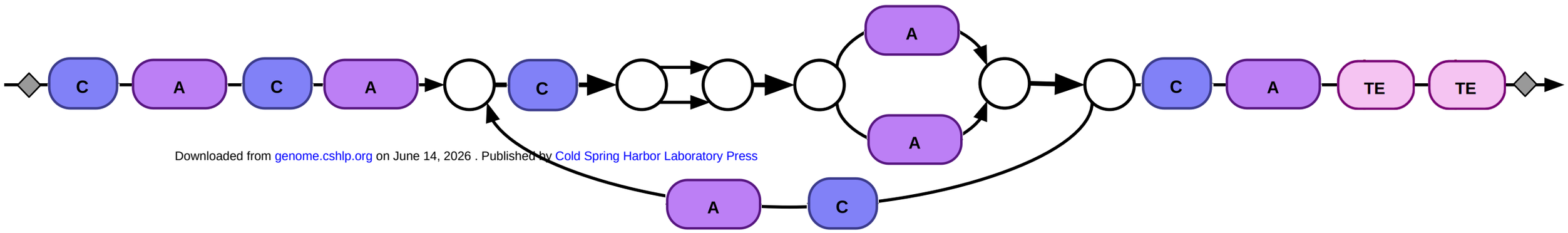

7
2
3
8
1
4
5,6


7
2
3
8
1
4
5
6

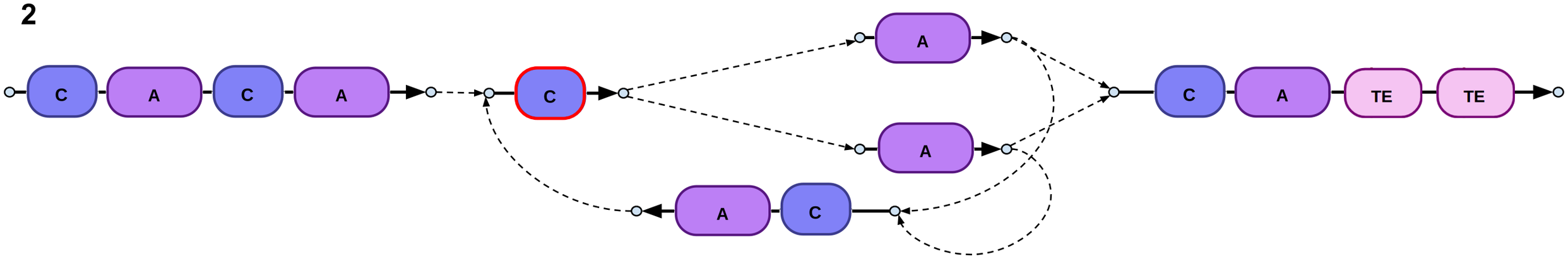

1
2
3
4
5
6
7
8


A
KS
AT
KR
C
A
A
C
A
C
KR
KR
C
A
A
KS
AT
KR
C
A
KR
C
A
KR
C
A
C
A

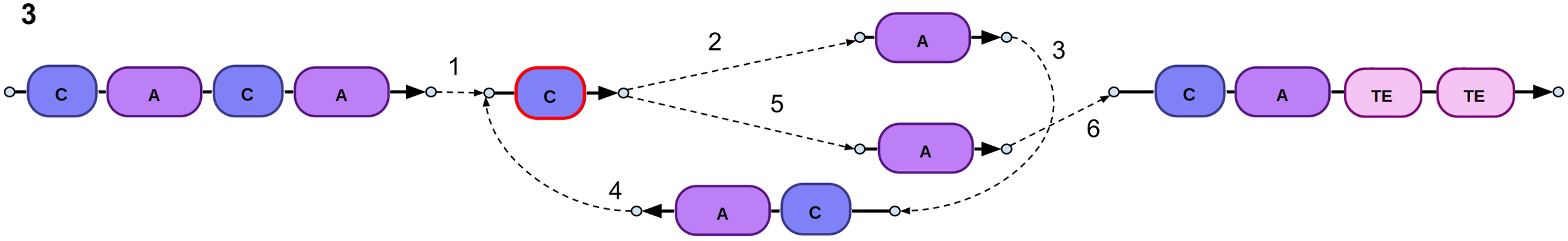
1



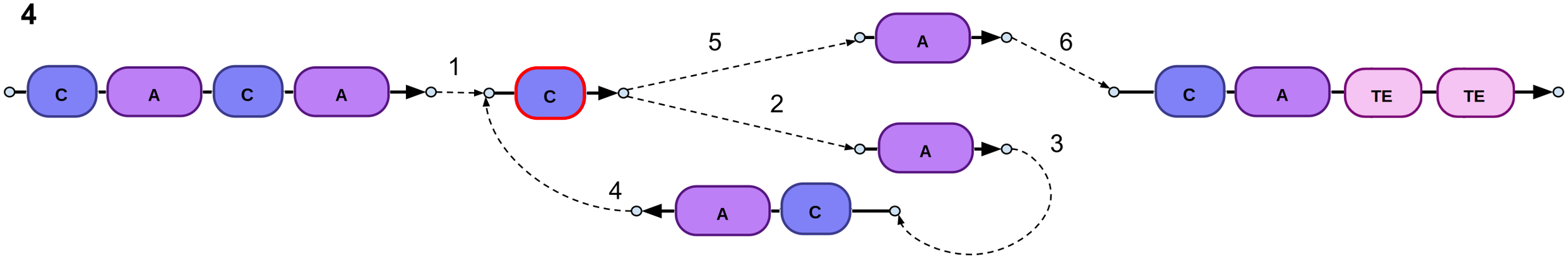
2



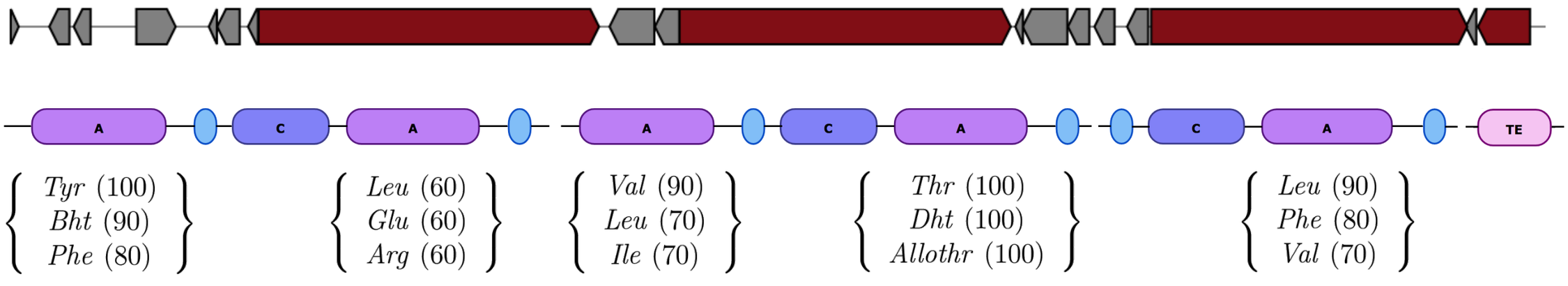
3



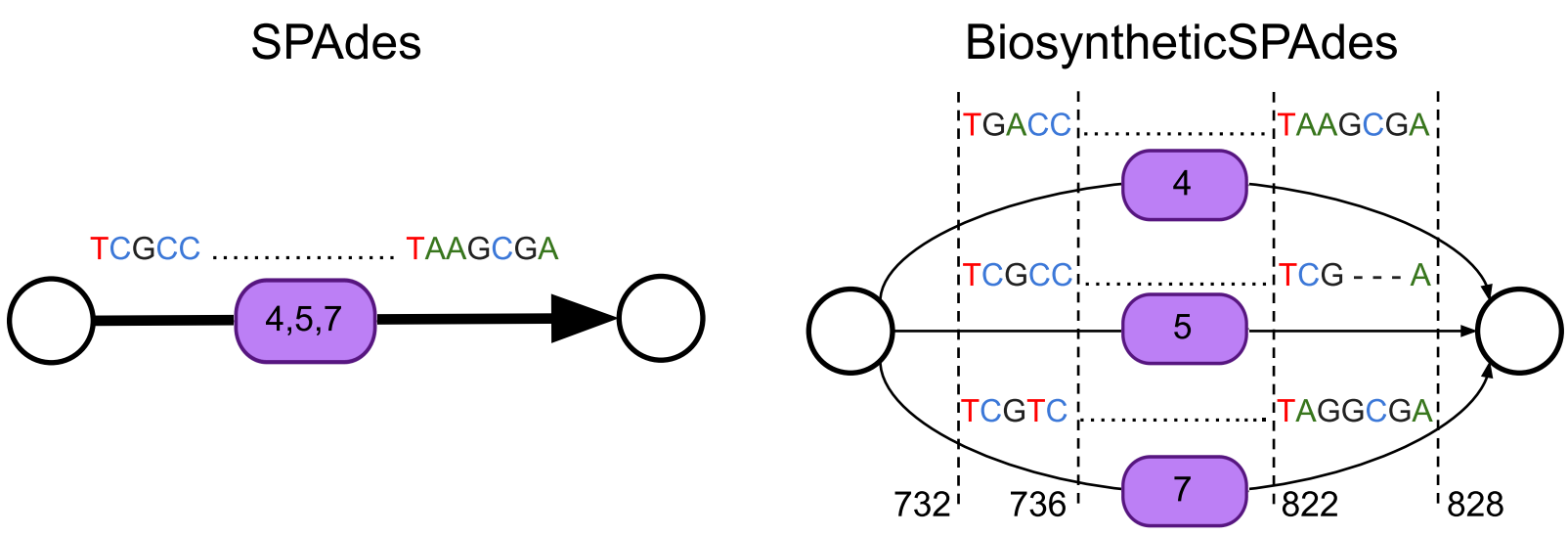
4

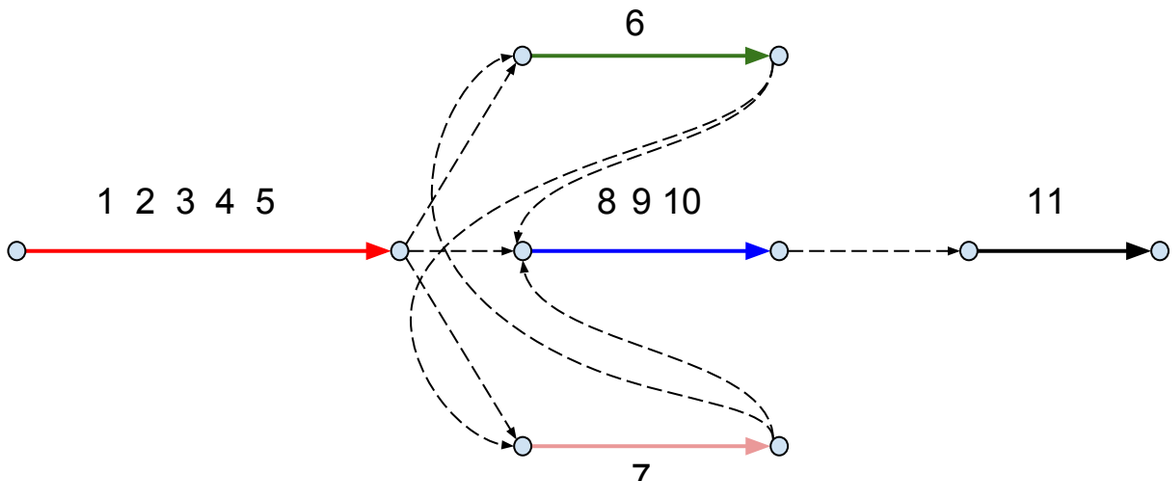


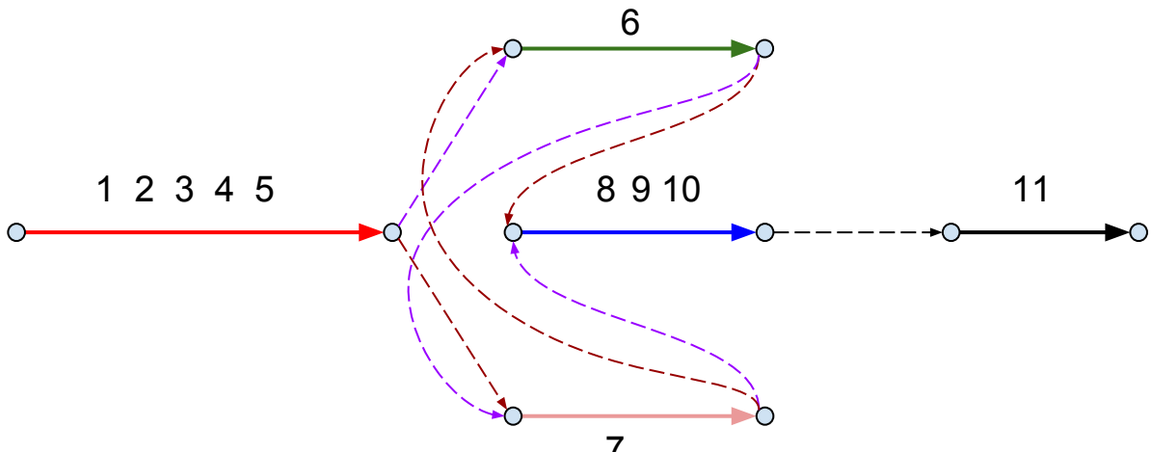
5



$\left\{ \begin{array}{c} Tyr\ (100) \\ Bht\ (90) \\ Phe\ (80) \end{array} \right\}$ $\left\{ \begin{array}{c} Leu\ (60) \\ Glu\ (60) \\ Arg\ (60) \end{array} \right\}$ $\left\{ \begin{array}{c} Val\ (90) \\ Leu\ (70) \\ Ile\ (70) \end{array} \right\}$ $\left\{ \begin{array}{c} Thr\ (100) \\ Dht\ (100) \\ Allothr\ (100) \end{array} \right\}$ $\left\{ \begin{array}{c} Leu\ (90) \\ Phe\ (80) \\ Val\ (70) \end{array} \right\}$


SPAdes
TCGCC ..... TAAGCGA
4,5,7
BiosyntheticSPAdes
TGACC ..... TAAGCGA
4
TCGCC ..... TCG - - - A
5
TCGTC ..... TAGGCGA
7
732
736
822
828


6
1 2 3 4 5
8 9 10
11
7


6
1 2 3 4 5
8 9 10
11
7

Assembler	Failed to assemble				
	all BGCs	BGCs with complexity 1-3	BGCs with complexity 4-6	BGCs with complexity 7-9	BGCs with complexity ≥10
SPAdes	149 (22%)	6 (2%)	23 (18%)	39 (37%)	81 (58%)
SPAdes + domain restoration	121 (18%)	9 (3%)	24 (19%)	30 (28%)	58 (41%)
biosyntheticSPAdes	69 (11%)	7 (2%)	16 (13%)	16 (16%)	30 (22%)
MEGAHIT	235 (35%)	`28 (10%)	34 (27%)	60 (58%)	111 (79%)
ABySS	227 (34%)	15 (5%)	35 (27%)	58 (56%)	117 (83%)
Total	665	293	128	104	140