THE NETHERLANDS
MAY 2017

WAGENINGEN
UNIVERSITY & RESEARCH

Systematic exploration of the Biosynthetic diversity of Glycopeptide Antibiotics

---

Bioinformatics Group

Diego Montiel
Supervisor: Dr. Marnix Medema
Examiner: Dr. Dick de Ridder

Master Thesis

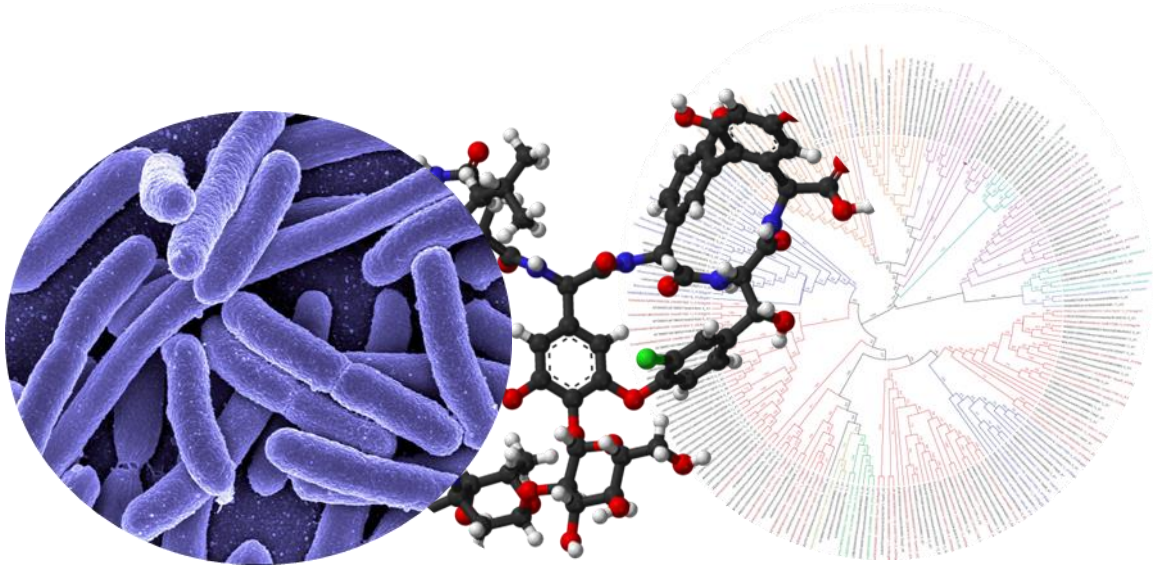

Table of contents

Abstract 3

Introduction 4

Materials and Methods 5

Gene cluster sequences 5

Extract Adenylation and Thiolation domain 5

Multiple Sequence Alignment and Phylogeny 5

Predicting substrate specificity 6

Calculating distance relationship between Gene Clusters 6

Results and Discussion 7

Exploration of Glycopeptide Biosynthetic Gene Clusters 7

Assigning substrates to Phylogeny 7

Architectural relationship 8

Predicting novel Glycopeptide Biosynthetic Gene Clusters 9

Assigning substrates to novel glycopeptide BGCs 9

Architectural relationship for novel glycopeptide BGCs 10

Absence presence of pattern key genes 13

Conclusions 14

References 15

Supplemental Information 17

Table 1: Glycopeptide and Novel Gene Clusters 17

Table 2: Best Models prediction with Prottest 2.4 Server 18

Figure 1: Workflow for the Identification of Novel GBGCs 18

Figure 2: Workflow project chart 19

ACKNOWLEDGEMENTS

I would first like to thank my thesis supervisor Dr. Marnix Medema of the Bioinformatics Group at Wageningen University and Research. He was a guide in this entire path and always willing to help and support into any trouble spot or question about my research or writing,. Secondly, I would also thank to Dr. Dick de Ridder as the examiner of this thesis, and I am gratefully indebted to his very valuable comments on this thesis.

I would also like to thank to the rest of members from the Bioinformatics Group, especially to Dr. Jorge Navarro who was involved and willing to help me during this research project. Also to all my friends who made all this process more enjoyable.

I would like to thank also to my sponsor, the Mexican institution CONACyT, without their financial support this work might not have been possible.

Last, but not least, the biggest acknowledgement goes for my mother Yara and my brother David. You both are my inspiration to keep moving forward and one of the main reasons of why I am here, this work goes especially for you two.

This accomplishment would not have been possible without you. Thank you!

Diego Fernando Montiel González

Abstract

Glycopeptide antibiotics are used in clinical treatment for skin or blood infections related to Gram positive bacteria[1]. But, due to antibiotic resistance, there is urgency for the discovery of novel glycopeptides that are capable of dealing with this. The discovery of novel glycopeptide antibiotics is possible by exploring their Biosynthetic diversity and machinery. This project consisted in a methodology to predict novel glycopeptide antibiotics through an automated pipeline that uses any gene cluster as input and then tries to find a match with well-known glycopeptides based on their domain annotation. Afterwards, using state of the art phylogeny, we constructed the architectural relationship with the alignment and distance matrix between both novel gene clusters and known glycopeptide antibiotics using their substrate specificity as criterion. We were able to predict 16 novel glycopeptide Biosynthetic gene clusters (BGCs) that showed to be related to several types of glycopeptide antibiotics such as vancomycin, complestatin and teicoplanin. Furthermore, we discovered a new diversity of glycopeptides that includes modification and/or different order in their residues. We observed similarities between the novel and known glycopeptides in heptapeptide backbone structure and domain annotation. Nevertheless, in order to prove this hypothesis, experimental characterization of the selected novel glycopeptide BGCs is needed in order to resolve the biological activity of its metabolic products.

Introduction

Glycopeptides are a class of non-ribosomal peptides (NRPs), microbial specialized metabolites with diverse chemical structures[2]. The biosynthetic pathways of glycopeptides are usually encoded by genes that are located closely in the genomes, i.e., biosynthetic gene clusters (BGCs). Since the mid 1950's glycopeptides have been a valuable resource of antibiotic compounds. For instance, vancomycin, and teicoplanin have played an important role in treating infectious diseases caused by Gram-positive bacteria[1], by blocking the steps in the biosynthesis of the peptidoglycan layer of bacterial cells walls. Both glycopeptides are natural products produced by actinomycete soil bacteria; vancomycin was the first glycopeptide isolated from *Amycolatopsis orientalis* and teicoplanin from *Actinoplanes teichomyceticus*. Furthermore, there has been an emergence of problematic resistant bacterial strains including Vancomycin-resistant enterococci (VRE), methicillin-resistant *Staphylococcus aureus* (MRSA) and Vancomycin-resistant *S. aureus* (VISA, hVISA, VRSA)[1]. This emphasizes the importance to find novel antibiotics with the potential to treat resistant Gram-positive infections[3]. Recently developed bioinformatics tools, such as antiSMASH[4] makes it feasible to predict NRPS biosynthetic gene clusters (BGCs) *in silico*. However, for most of these identified BGCs, the metabolic products still remain unknown.

Regarding the synthesis mechanism of NRPs, adenylation (A) domains are known to be responsible for each adding one amino acid to the natural product scaffold through the formation of an aminoacyl adenylate via covalent bonding of the activated amino acid as a thioester to the 4'-phosphopantetheinyl (4'PPant) cofactor of the thiolation (T) domain[5]. Previous studies revealed that substrate specificity is encoded in the adenylation domains sequences, i.e., evolutionarily closely related adenylation domains can share the same substrate specificity[6]. This indicated that one can infer the substrate specificity of adenylation domains in a glycopeptide BGC via phylogenetic analysis, which can shed light on revealing the structure of the metabolic product.

Before we solve the question of how to infer the chemical structure of novel glycopeptides from the BGCs sequence information, we first need to figure out what are the evolutionary relationships between known glycopeptide BGCs by using their assembly lines (substrates). To do that, we performed a phylogenetic analysis of A-domains from 11 glycopeptide BGCs and identified their functional clades (monophyletic group) in which A-domains share the same substrate specificity. Also, we mapped the absence/presence pattern of genes involved in the biosynthesis of their metabolic precursors. Then, this information was integrated to infer the evolutionary distance between these glycopeptide BGCs and group them into subfamilies[7].

Next, with our methodology we carried-out the aforementioned analysis by adding 16 novel glycopeptide BGCs from our Glycopeptide-Prediction pipeline. As such, we could figure out in the phylogenetic analysis i) whether any novel BGCs are evolutionary closely related to certain well-characterized glycopeptide BGCs by inspecting manually in each of their monophyletic groups, ii) the diversity between known and novel glycopeptides, showing how conserved certain modules are on each of the assembly lines; this also revealed the taxonomy distribution among glycopeptides, iii) the predicted moieties of the novel glycopeptides are well supported by the presence of the genes involved in the predicted precursor synthesis.

Materials and Methods

Gene cluster sequences

The glycopeptide BGC protein sequences and GenBank entries used here are available in the NCBI website and in the Minimum Information about a Biosynthetic Gene Cluster (MIBiG) database (Supplementary Table 1). For the prediction of novel glycopeptide BGCs, we used as input antiSMASH results from the GenBank database from 2015 and we also looked for GCs homologs by using the X domain of the known glycopeptides with a non-redundant (nr) database search with BLAST[8]. The reason of why we used the X domain sequence as our query in the BLAST search is because this domain is known to be a well conserved domain exclusively from glycopeptides that has comes from remnants of A-domains during evolution[5].

Then, we implemented certain rules to validate if a certain GC may encode the biosynthetic pathway towards a glycopeptide using an in-house Python script and for profile domain match we used hmmsearch[9]. In hmmsearch we used as query all profile Hidden Markov Model (pHMM) that belongs to the NRPs class from the Pfam database and as database all GCs that were obtained from antiSMASH results and from the previous BLAST search. Then, with our pipeline we iterated through the hmmsearch output looking for GCs in which their domain annotations contains more than four NRPS adenylation domains that together have multiple condensation (C) domains for which the Cglyc domain (glycopeptide condensation domain)[10] and X domain (inactive C domain at the C-terminus)[5], both exclusively from glycopeptides, gave a higher score with hmmsearch than the other Condensation domains subtypes (Condensation_LCL, Condensation_DCL, Condensation_starter, Condensation_Dual)[5], otherwise it may result in several hits of different profiles for the same sequence. They also needed to contain at least one Epimerization domain which also should score higher than the conventional Condensation domains on the same part of the protein sequence[1].

Extract Adenylation and Thiolation domain

The A-domain in order to recognize a specific amino acid and activates it, it needs to be follow of a covalent bonding Peptidyl carrier protein, also called T-domain[5]. So, in this sense we extracted both A-domain and T-domain sequences from the gene clusters coding sequence (CDS) with the bioinformatics tool hmmsearch along with a self-written Python script. Then, in order to exclude A-domains that are not following by a T-domain we filtered A-domains sequences that were followed by a T-domain and convert them into fasta format.

Multiple Sequence Alignment and Phylogeny

All A-domain sequences from identified BGCs were automatically aligned with MUSCLE v3.8.31[11] with default parameters[12]. The most appropriate model of protein evolution for the phylogeny was selected using the ProtTest v2.4 Server[13] (Supplementary Table 2). ProtTest used as selection the smallest AIC (Akalke Information Criterion) among different models and parameters[14]. For the construction of the phylogenetic tree from the multiple sequence alignment (MSA) of A domains sequences we used RaxML[15] for a Maximum Likelihood method with a Bootstrap of 100; the employed substitution matrix (ProtTest) was the WAG matrix model with I distribution and Gamma distributed rate variation.

Predicting substrate specificity

The basic structure of domain annotations among glycopeptides needs to have at least but not exclusively C, A and T-domains, ending with a TE-domain. According to literature, it is well known that A-domains are responsible for the substrate recognition and activation and consequently their sequence is used to predict the amino acid encoded by the module via covalent bonding of the T-domain[1] while C-domain catalyzes the elongation reaction of the Peptidyl chain.

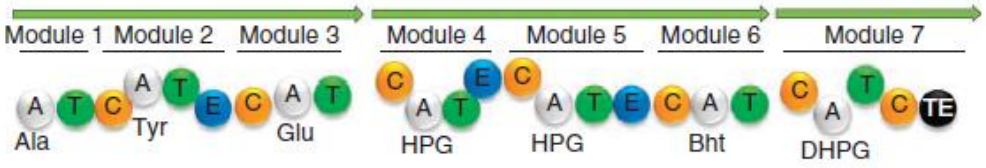

Figure 1: Gene cluster for pekiskomycin. A module represents the group of domains involved in the incorporation of one amino acid into the growing peptide chain. We see that for the substrate formation is needed the subsequent modules of A-T-C[5].

In order to correctly assign the substrate specificity of A-domain sequences from known BGCs, we dived into literature for each case to identify their substrate specificities. In the case of the novel BGCs, their substrate specificities A-domains were predicted by manually looking at the closer cluster from each substrate on the monophyletic groups.

Calculating distance relationship between Gene Clusters

To gain an overview of the evolutionary relationships between the GCs we performed a phylogenetic analysis of the A-domain specificities present in 27 glycopeptide BGCs and several subgroups were identified (see results). For this tree, we constructed pseudo-sequences from A-domain substrate specificities for all BGCs representing the functional architectures of the encoded assembly lines to estimate the evolutionary distances between the Gene Clusters (Figures 3 and 5). We used a distance metric adapted from the BiG-SCAPE software (https://git.wur.nl/medema-group/BiG-SCAPE), with domain types defined either according to their reference literature or to the phylogenetic prediction tree (Figures 2 and 4) with weights of the Jaccard index, Goodman–Kruskal (GK) Index and domain duplication (DDS) index of 0.5, 0.25 and 0.25 respectively[6]. The Jaccard Index measures the amount of shared domains between a pair of BGCs, the DDS Index that stands for Duplicate Domain Score where calculates the difference in domain sequence, and the GK Index measures the synteny between two lists of domains.

Code availability of all data and Python scripts used for prediction of Glycopeptide Biosynthetic Gene Clusters are available at: https://github.com/dmontielg/Glycopeptides

Results and Discussion

Exploration of Glycopeptide Biosynthetic Gene Clusters

Assigning substrates to Phylogeny

To explore the Biosynthetic diversity of glycopeptides antibiotics we constructed a phylogeny with the Multiple Sequence Alignment (MSA) from the A-domains of 11 well-known glycopeptide BGCs (figure 2). These A-domains correspond to each substrate specificity prediction that we manually assigned according to literature research. Among these substrates we identified three nonproteinogenic amino acids; i) 4-hydroxyphenylglycine (Hpg)[16], ii) 3,5-dihydroxyphenylglycine (Dpg)[17], iii) β-Hydroxytyrosine (Bht)[18] and six proteinogenic amino acids; i) leucine (Leu), ii) tyrosine (Tyr), iii) asparagine (Asn), iv) tryptophan (Trp), v) alanine (Ala), and vi) glutamic acid (Glu).

The Maximum Likelihood[15] phylogeny showed that complestatin A-domains sequences often clustered outside of the corresponding substrate clade. The phylogeny also shows that the A-domain sequences specificities for the non-proteinogenic amino acids 4-hydroxyphenylglycine (Hpg) and 3,5-dihydroxyphenylglycine (Dpg), are clustered together, as is the case for the non-proteinogenic amino acid Bht. We will further discuss which characteristics these amino acids have and how they are related to other proteinogenic amino acids as well.

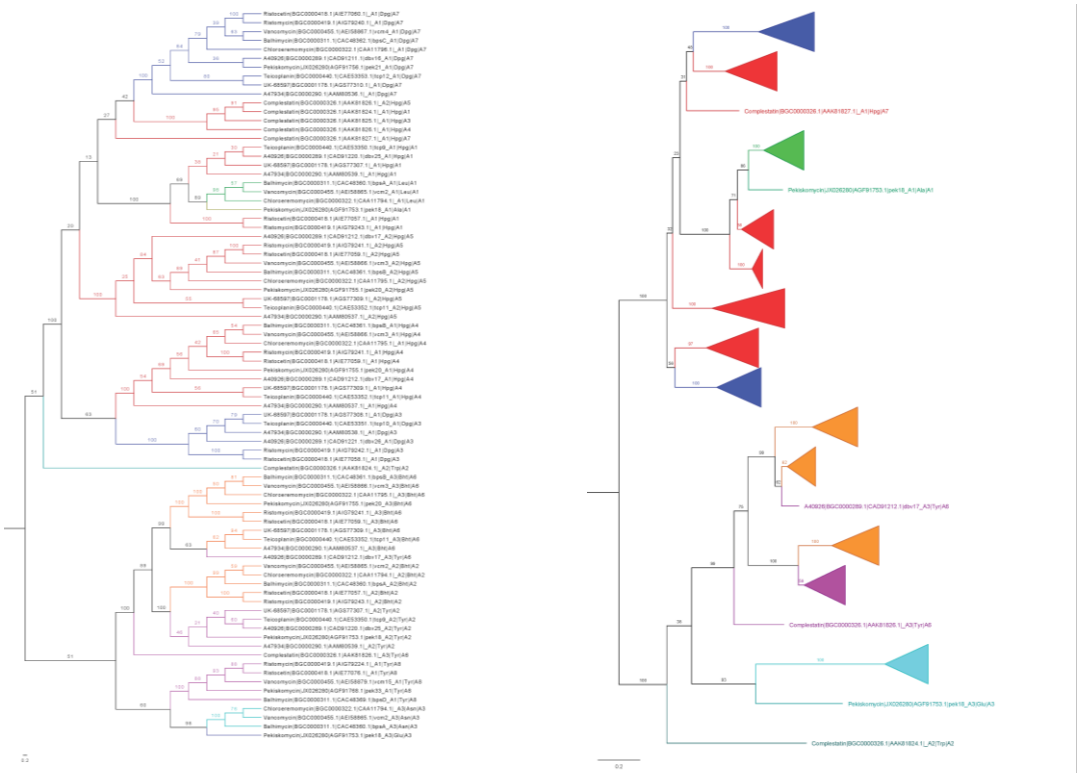

Figure 2: MSA from A domains of 11 known glycopeptide BGCs. Non-proteinogenic amino acids: red-Hpg, blue-Dpg, and orange-Bht. Proteinogenic amino acids: green-Leu, light-green-Ala, pink-Tyr, sky-blue-Asn, and others-Trp-Glu.

We identified the monophyletic clades with shared substrate specificity, and observed how these substrates form the assembly line on each gene cluster. One of the monophyletic clades showed glutamic acid (Glu), asparagine (Asn), and tyrosine (Tyr) as proteinogenic amino acids and Bht as nonproteinogenic amino acid; this is in concordance with this monophyletic group due to it belonging to balhimycin, chloroeremomycin and vancomycin. These glycopeptide BGCs share identical heptapeptide backbones, where we can see Leu and Asn are present on the first and third module (figure 3). We also found that that the substrate for glutamic acid belongs to the unusual glycopeptide Pekiskomycin. This is a rare glycopeptide where Ala and Glu are exclusively present in the first and third module respectively. Despite Bht having one extra hydroxyl group compare to Tyr, they maintain chemical similarities that are reflected by their A domains belonging to the sister clades.

Then, the monophyletic clade identified with the substrate Hpg and Dpg as nonproteinogenic amino acids we can observe that due to N-methylation Ala the first substrate that encodes for Leu of the BGCs balhimycin, chloroeremomycin and vancomycin[1]. Pekiskomycin is a rare type of glycopeptide antibiotic whose first substrate Ala showed to be closer related to Leu, and seems to be part of the same sister group.

Modification in the heptapeptide backbone residue for each glycopeptide antibiotic is caused by specific tailoring enzymes such as glycosyltransferases (Gtfs), acyltransferases (Atfs), halogenases, sulfotransferases (Stfs) and methyltransferases (Mtfs)[19]. For instance, in the case of vancomycin, balhimycin and chloroeremomycin, methylation occurs at the N-terminal amine of the heptapeptide, specifically on the N-terminal Leu. Pekiskomycin, on the other hand, known for its unusual heptapeptide structure, encodes only one predicted N-Mtf in the BGC[1], even though its substrate modification is due to dimethylation at the N-terminal Ala.

Architectural relationship

Subsequently, we calculated the evolutionary distance by considering the A-domain functional composition and sequence similarity according to the predicted substrates, and applying the UPGMA algorithm on the distance matrix to identify glycopeptide subfamilies (Figure 3).

We obtained an architectural relationship tree among the 11 BGCs showing the similarities and differences between the assembly lines. As a first instance we can observe that chloroeremomycin, vancomycin and balhimycin are conserved on the entire assembly line and are the only ones who share on modules one and three Leu and Asn respectively. While the rest of glycopeptides architecture showing at the same positions to have Hpg and Dpg instead on the first and third module, only excluding complestatin where it has double Hpg in these first and third module and pekiskomycin with Ala and Glu in the same modules.

Complestatin and pekiskomycin seemed to have unique heptapeptide structure. In case of complestatin this is because it is composed exclusively of one type of non-proteinogenic amino acid (Hpg) in the assembly line, and two proteinogenic amino acids, tryptophan (Trp) and Tyr. The A domain sequences from complestatin did not cluster in a specific clade; for example, the A domain that encodes for Trp is obviously different from the rest of proteinogenic amino acids due to the largest aromatic ring and hydrophobicity. In the same context, pekiskomycin showed to be the only glycopeptide having Ala and Glu. These unique properties make them different and they do not cluster easily with other glycopeptides.

Ristomycin A and Ristocetin showed to have similar heptapeptide structures, each having exclusively nonproteinogenic amino acids from *Amycolatopsis japonica,* and *Amycolatopsis lurida* respectively. We can also see that A47934, teicoplanin and UK-68597 are in the same monophyletic clade, this is in agreement with the fact that they are often called as "teicoplanin type"[1], because they possess the same heptapeptide architecture. This also applies for "vancomycin type" in vancomycin, chloroeremomycin and balhimycin, as we can observe this is why they belong to the same monophyletic group (figure 3).

The only difference between A40926 and the teicoplanin type is the lack of hydroxyl group in the sixth module, making this a Tyr instead of Bht. Nevertheless, these glycopeptide GCs maintained in the cluster group in the phylogeny. This analysis showed that there are different monophyletic clades, probably cause by the different substrates in the assembly line. It seemed to be grouped in pekiskomycin, complestatin[20], vancomycin-type[21], and teicoplanin-type[22]. This suggests that multiple events might have taken place during the evolution of this BGC family and due to tailoring enzymes leading to modification in specific substrates.

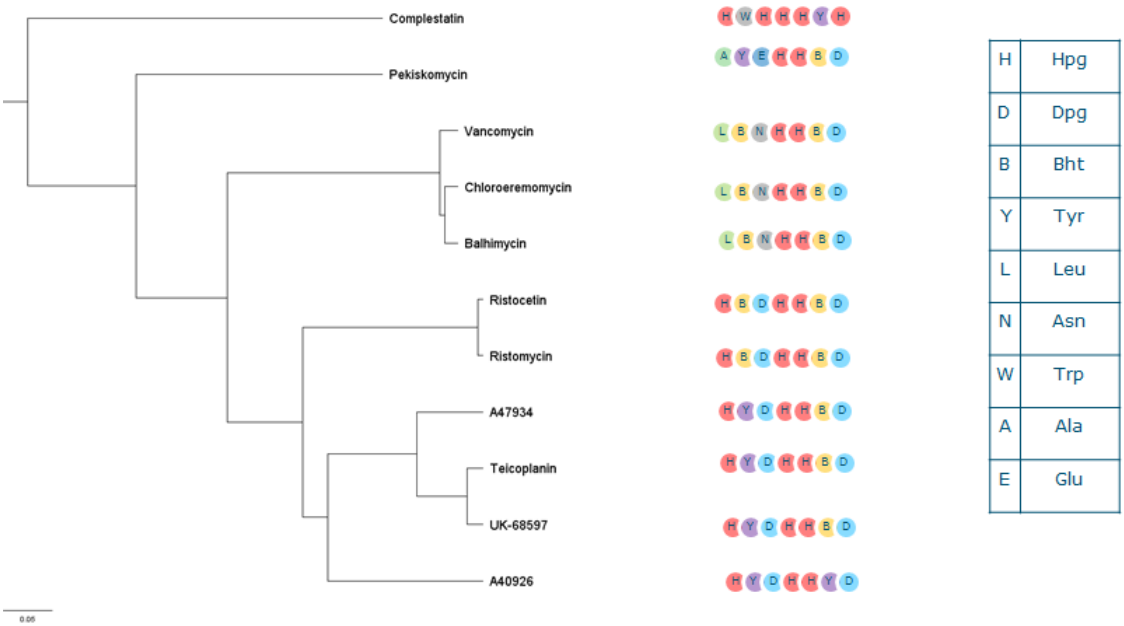

Figure 3 Architectural relationships based on the distance similarities and pseudo-sequences between 11 known glycopeptides BGCs.

Predicting novel Glycopeptide Biosynthetic Gene Clusters

Assigning substrates to novel glycopeptide BGCs

In order to predict novel glycopeptide BGCs we gathered BGC prediction from antiSMASH (see methods) and we manually looked for homolog of the X-domains with a BLAST search from the 11 known glycopeptides. We used antiSMASH results and the homologs found as input in our pipeline (see methods) to find potential novel glycopeptide BGCs. We were able to find 16 potentially novels glycopeptide BGCs. Using the rules mentioned before in methods, It might be possible that we have left some novel glycopeptides BGCs behind probably because there might still be other, untapped criteria to use besides the Cglyc, X or E domain. (See methods). Nevertheless, we can be sure that the prediction from our pipeline detected in all the cases gene clusters which showed similar backbone structure to

glycopeptides BGCs. There were three gene clusters found that are lacking of one A-domain, which we presumed should be at the beginning of the first module.

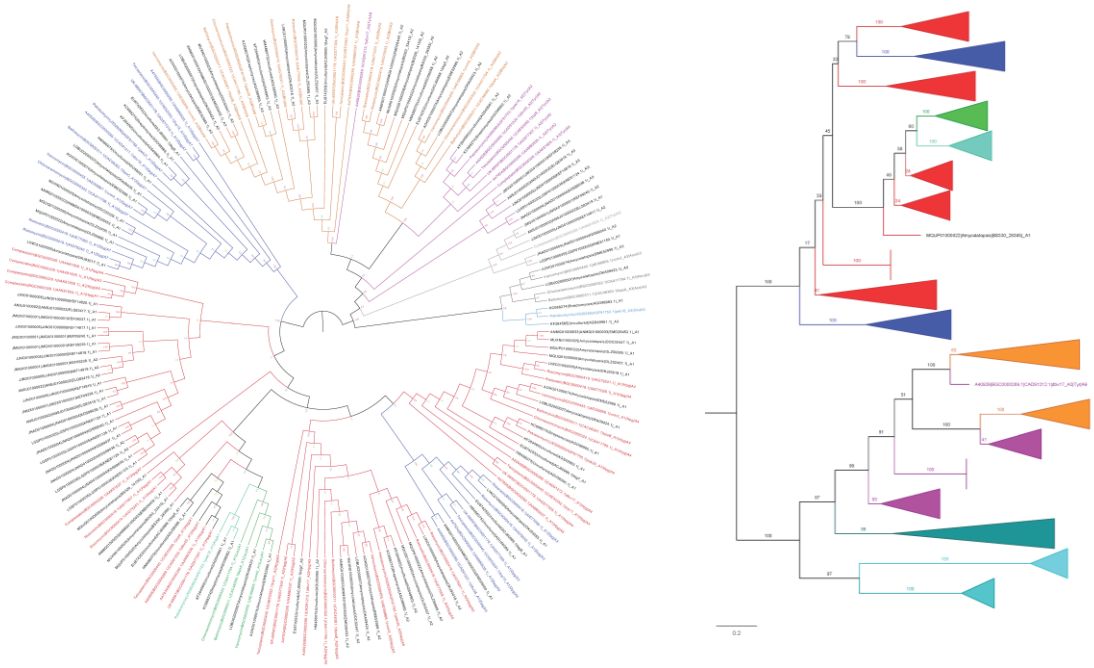

Figure 4: MSA of 202 Adenylation domain sequences from known and novel GBGC. Non-proteinogenic amino acids: red-Hpg, blue-Dpg, and orange-Bht. Proteinogenic amino acids: green-Leu, light-green-Ala, pink-Tyr, sky-blue-Asn, and others-Trp-Glu.

We aligned 202 A domain sequences corresponding to known and novel glycopeptide GCs. After re-doing the previous approach with our designed pipeline we constructed a new phylogeny (figure 4) discovering and predicting the substrates according to the monophyletic clades. We also used NRPS predictor2, a web based tool for A domain prediction based on Transductive Support Vector Machines (TSVMs) [27] to compare the prediction results, and to try to resolve few cases of discrepancy where the monophyletic group offered a less-than clear view for the assignment of specificities. With this new phylogeny we found that most of the A domain prediction matched with the results from NRPS predictor2, excepting the case of pekiskomycin. For this, NRPS predictor2 was not able to predict most of their A domains because this glycopeptide it is not yet in the training set of this web tool.

Architectural relationship for novel glycopeptide BGCs

The evolutionary distance of the glycopeptide BGCs architecture was calculate by adding the novel glycopeptide BGCs and the phylogeny of the architectural relationship is based on the 26 glycopeptide BGCs was constructed (figure 5). Then, we visualized the module predictions for each of the assembly lines. The GCs with black labels represent the predicted glycopeptide assembly lines and how they cluster in the tree with the known glycopeptides GCs (blue label). The gene clusters with accession number AF386507[20], AJ223999[23], and AJ632270[22] were found to correspond to the known

glycopeptides complestatin, chloroeremomycin and teicoplanin respectively. They also have been identified as glycopeptide GCs in literature. Furthermore, by running an antiSMASH prediction in all the Novel Gene Clusters showed that some are fragmented such as the case of; LGSP01000035, ANSJ01000022, and LIWC01000005, where the substrate at the beginning, of the first module is missing. So, due to the lack of A-domain this substrate cannot be predicted.

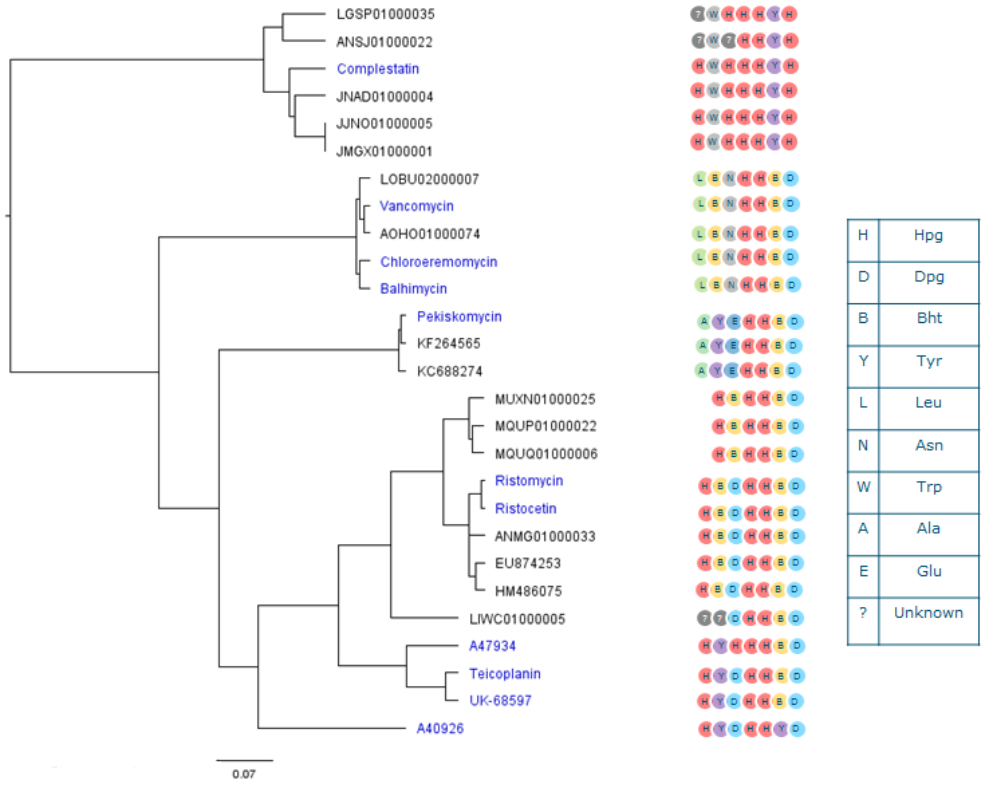

Figure 5: Architectural relationship between 11 and 16 novels GBGCs. Phylogeny of the calculated distance between the gene clusters. Highlight in blue belong to the known GBGC.

Additionally, GBGCs with the accession number LOBU02000007, ANMG01000033, MUXN01000025, MQUP01000022 and MQUQ01000006 are quite recent (2016 and 2017) gene clusters from Whole Genome Shotgun Sequence whose papers unfortunately are still unpublished at this time. So, there is not further information available from these. However, regarding substrate prediction of LOBU02000007, and ANMG01000033, they present similar structures to vancomycin and ristomycin A respectively. The ones identified as MUXN01000025, MQUP01000022 and MQUQ01000006 seem to cluster together in an independent clade but still have to ristomycin A and ristocetin as sister group. In figure 5 there are three GCs (MUXN01000025, MQUP01000022 and MQUQ01000006) that belong to genus *Amycolatopsis* (Supplementary Table 1) and were found as mis-annotated due to the fact that they are missing a module, probably at the beginning of the GC module in order to be hepatapetides[24]. One possible reason to this is because they are recent GCs (January 2017) and their papers are still unpublished on the NCBI website there may still be missing some annotation.

Gene Clusters LGSP01000035, ANSJ01000022 belong to *Streptomyces fradiae* organism and JNAD01000004, JJNO01000005 and JMGX01000001 to *Streptomyces rimosus* (Supplementary Table 1). These organisms are already used for different types of antibiotics or as inhibitors of ATP synthase such as oligomycin[25], oxytetracycline and tylosin[26]. In our method we observed (figure 5) that these predicted novel glycopeptide BGCs show similarities in backbone structure with Complestatin, which it makes sense due to complestatin belongs to *Streptomyces lavendula.*

AOHO01000074[27] is a draft genome Gene Cluster sequence from *Amycolatopsis decaplanina*. Despite the target of this contig not being specified in literature, we can presume that is closely related to vancomycin. Due to both GCs AOHO01000074 and vancomycin belong to the genus *Amycolatopsis* this gave us some certainty that this novel GCs is a potential candidate for glycopeptide biosynthesis.

There are some novels GCs that according to literature they have already been predicted as glycopeptides, but don´t seem well characterized and have also not been uploaded to MIBiG database. Nevertheless, these predictions are: EU874253, HM486075, KF264565, and KC688274.

KF264565[28] has a literature reference from a pipeline analysis of uncharacterized glycopeptides. But we identified it in the KC688274 and pekiskomycin cluster group. Pekiskomycin is known to have unusual specificities due to o-methylation by having Ala and Glu at positions one and three. We can observe the same situation in the two novel GCs. There are no publications about KC688274.

The EU874253[29] GCs encodes identical biosynthesis and heptapeptide to HM486075[30], ristomycin A and ristocetin. Even though EU874253 comes from the TEG (teicoplanin-like environmental DNA derived gene cluster) pathway[1], the difference of this with teicoplanin is the presence of Tyr instead of Bht in the second module in the assembly line[29]. The authors, who reported to the gene cluster EU874253, extracted eDNA (environmental DNA) from several soil samples and looked for the presence of OxyC, an enzyme whose function is to catalyze the formation of the C-C bond between positions five and even of teicoplanin. As we observed in the assembly line these two positions remain well conserved in glycopeptide GCs for vancomycin-type and teicoplanin-type.

Moreover, the authors who discovered the gene cluster HM486075 used a culture-independent strategy from soil environment with eDNA that contained collections of tailoring enzymes, in particular Sfts. In their experiment they took a sample of six glycopeptide BGCs that were isolated from soil eDNA, and found that this enzyme appeared with more frequency in glycopeptides[30].

For the case of KF264565[28], it was also discovered from eDNA taken from soil samples where they applied barcorded PCR amplicons to identify sequences bearing close resemblance to biosynthetically gene clusters. Subsequently, these amplicons are mapped onto arrayed metagenomic libraries to guide the recovery of targeted gene clusters using as target sequences A and PKS ketosynthase (KS) domains.

The gene cluster KC688274, with strain sp. “WAC4229”, has already been identified to have strong similarity with the glycopeptide pekiskomycin. The authors developed an approach for the discovery of new antibacterials from *Actinomycetes* using resistance as a trait for selecting glycopeptide-producing organisms and by applying a screen for resistance they were able to isolate pekiskomycin[31].

Absence presence of pattern key genes

In order to support our predictions of non-proteinogenic specificities, from previous phylogeny (Figures 4 and 5), we mapped the absence and presence of pattern key genes involved in the biosynthesis of the metabolic precursors in the GCs. First, according to literature we identified the genes involved in the biosynthesis pathway of the non-proteinogenic amino acids, with the accession numbers i)Hpg[16]: CA11792 (Pdh), CAA11761 (HmaS), CAA11762 (Hmo), ii) Dpg[17]: CAC48378 (DpgA), CAC48379 (DpgB), CAC48380 (DpgD), CAC48381 (DpgD), for both Hpg and Dpg: CAA11790 (HpgT) and iii) Bht[18]: CAC48368 (BpsD), CAC48369 (OxyD), and CAC48370 (Bhp). Then, we performed a BLAST search with these genes as query against the 27 GCs to generate (Table 1) where the green blocks show presence of the key gene precursor in the GC and the white blocks represent absence.

The GCs labeled in black belong to the known glycopeptides. We can observe that complestatin lacks the gene precursors for Dpg except one and all the genes for Bht biosynthesis. This was expected due to complestatin only possessing Hpg as non-proteinogenic amino acid. A40926 did not present the precursors of Bht due to it not being present as the final product in the assembly line. Chloreomomycin, on the other hand, showed to be fragmented into two clusters, so we performed a BLAST against the two clusters present there in order to find hits with the gene precursors. Teicoplanin, A47934 and UK-68,597 for some reason don´t seemed to have significant hits with Bht genes, but after a close inspection in the GenBank file, the putative hits we found correspond to regions annotated to involved in the production of modules four, five and six, and module six stands for Bht. In case of the rest of glycopeptides GCs they present all the precursors; this showed concordance with the architectural relationship (Figure 5). Furthermore, from Table 1, the GCs labeled in color green it belong to the novel glycopeptide BGCs that according to the architectural relationship and the prediction from Figure 3 present all the necessary gene precursors that are needed for the biosynthesis of the non-proteinogenic amino acids.

Finally, the GCs with names in red curiously showed a lack of genes involved in the biosynthesis of amino acids which, according to our prediction should be present. A possible explanation could be that the gene clusters are not complete or that these key genes are found somewhere else in the genome. Such is the case for ANMG01000033 which lacked most of the precursors showing only presence of Pdh, the first step in the production of Hpg. EU874253 lacked of one precursor for Hpg and Bht and the three genes in charge of Dpg. KF264565 and KC688274 that are closely similar to pekiskomycin both seemed to lack DpgD.

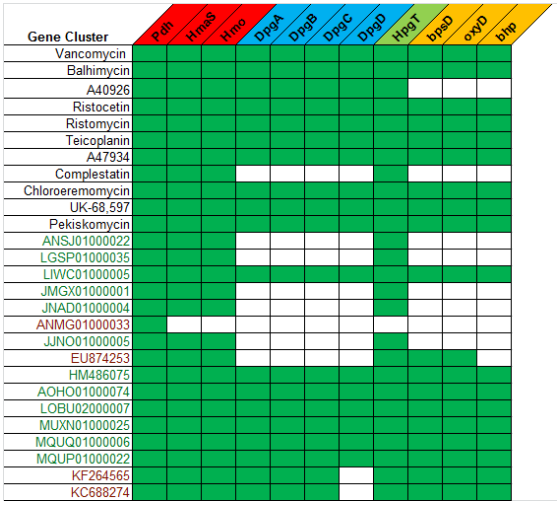

Table 1: Matrix table containing map of the absence/absence patterns of key genes involved in the biosynthesis of metabolic precursors.

Conclusions

We propose a new methodology for finding novel glycopeptide GCs, along with a different and novel phylogenetic approach for assigning substrate specificity. Furthermore, this project gives to biologists and researches further information about the biosynthetic diversity of glycopeptide antibiotics and pattern key gene precursors to be used in molecular engineering[32]. Our approach *in silico* was able to detect strong similarities with glycopeptides from novel GCs based in their domain annotation. We defined certain rules that allow us to compare from each GC it´s domain profile. This method could be improved by taking in account more factors such as different enzymes or other domain criteria's. As a next step this work may lead to the design of experiments to experimentally characterize the selected novel glycopeptide BGC and resolve the structure and biological activity of its metabolic products.

Vancomycin as the first glycopeptide has shown us that the evolution of their substrates has been widely conserved in the three nonproteinogenic amino acids Bht, Dpg and Hpg where appear to be conserved in "vancomycin-type" and "teicoplanin-type" GCs on modules four, five and six. On the other hand, the first modules showed to be different between types of glycopeptide such as teicoplanin, and vancomycin. There is not a ground truth about how the assembly line in glycopeptides should be, such as the case of complestatin where we can see that the only substrates Hpg-Asn-Trp are present here. The taxonomy distribution shows that in most of the cases organisms of the same genus cluster closely together. *Amycolatopsis* is the widest organism used in glycopeptides, but there are two cases where two identified glycopeptides were taken from a metagenomics sample in soil bacterium[29]. *Streptomyces* has shown to be the only organism used for complestatin. A40926 showed to be different at the beginning on modules one and two, by having two subsequent Hpg, the genus *Nonomuraea* is the only case where is present.

During recent years work has been done for the discovery of glycopeptides. Usually researches refer to glycopeptides into two types; vancomycin type and teicoplanin type because they showed to have well conserved modules among the diversity of glycopeptides. Also the taxonomic distribution of glycopeptides has proven that *Amycolatopsis* is the mostly organisms for glycopeptides. Also, researches have used metagenomics in soil from uncultured bacteria in order to discover unusual glycopeptides, just is the case of pekiskomycin. Even though the machinery of these biosynthetic gene clusters has been conserved with Hpg, Dpg and Bht from non-proteinogenic amino acids we don't disregard that there might be more combinations and other tailoring enzymes of these NRP that might allow us to find better and more resistant antibiotics.

Metagenomics have been used in recent years for the discovery of new compounds and natural products like antibiotics by extracting eDNA from soil, such as using enzymes like OxyC, tailoring enzymes like Sfts, barcoded PCR amplicons and resistant screening. It may be possible to merge all these approaches in metagenomics to do both *in silico* and experimental work. This work envelope the systematic exploration of glycopeptides based on their domain annotation and A-domain prediction. It could be possible to complement all this work to explore the glycopeptide diversity. This could lead to Novel GBGCs that may bear to similar biosynthesis pathways.

Glycopeptides are well known to be effective as a last resource for treating blood and skin infections for example. Since antibiotics were developed, bacteria, has proven to be a formidable rival, because it has been able to evolve and adapt to become resistant to this medications. In that sense we need to work harder in order to understand the reason on how they are able to fail to be active against bacteria

together with the discovery of novel enzyme activities is crucial to the development of novel antibiotics[33]. This thesis project revealed some evolutionary genomic engineering principles underlying glycopeptide molecular diversity, which may pave the way for future discovery and engineering of novel glycopeptide antibiotics.

References

1. Yim, G., Thaker, M. N., Koteva, K. & Wright, G. Glycopeptide antibiotic biosynthesis. *J. Antibiot. (Tokyo).* **67,** 31–41 (2013).

2. Fischbach, M. A. & Walsh, C. T. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: Logic machinery, and mechanisms. *Chemical Reviews* **106,** 3468–3496 (2006).

3. Yim, G. *et al.* Harnessing the synthetic capabilities of glycopeptide antibiotic tailoring enzymes: Characterization of the UK-68, 597 biosynthetic cluster. *ChemBioChem* **15,** 2613–2623 (2014).

4. Weber, T. *et al.* AntiSMASH 3.0-A comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.* **43,** W237–W243 (2015).

5. Rausch, C., Hoof, I., Weber, T., Wohlleben, W. & Huson, D. H. Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evol. Biol.* **7,** 78 (2007).

6. Nguyen, D. D. *et al.* Indexing the Pseudomonas specialized metabolome enabled the discovery of poaeamide B and the bananamides. *Nat. Microbiol.* **2,** 16197 (2016).

7. Cimermancic, P. *et al.* Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158,** 412–421 (2014).

8. Johnson, M. *et al.* NCBI BLAST: a better web interface. *Nucleic Acids Res.* **36,** (2008).

9. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* **39,** (2011).

10. Medema, M. H. *et al.* AntiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.* **39,** (2011).

11. Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).

12. Okonechnikov, K. *et al.* Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics* **28,** 1166–1167 (2012).

13. Abascal, F., Zardoya, R. & Posada, D. ProtTest Manual. *Bioinformatics* **21,** 2104–2105 (2005).

14. Posada, D. & Buckley, T. R. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst. Biol.* **53,** 793–808 (2004).

15. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large

phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).

16. Hubbard, B. K., Thomas, M. G. & Walsh, C. T. Biosynthesis of L-p-hydroxyphenylglycine, a non-proteinogenic amino acid constituent of peptide antibiotics. *Chem. Biol.* **7,** 931–942 (2000).

17. Chen, H., Tseng, C. C., Hubbard, B. K. & Walsh, C. T. Glycopeptide antibiotic biosynthesis: enzymatic assembly of the dedicated amino acid monomer (S)-3,5-dihydroxyphenylglycine. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 14901–14906 (2001).

18. Puk, O. *et al.* Biosynthesis of chloro-??-hydroxytyrosine, a nonproteinogenic amino acid of the peptidic backbone of glycopeptide antibiotics. *J. Bacteriol.* **186,** 6093–6100 (2004).

19. Lamb, S. S., Patel, T., Koteva, K. P. & Wright, G. D. Biosynthesis of sulfated glycopeptide antibiotics by using the sulfotransferase StaL. *Chem. Biol.* **13,** 171–181 (2006).

20. Chiu, H. T. *et al.* Molecular cloning and sequence analysis of the complestatin biosynthetic gene cluster. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 8548–53 (2001).

21. Xu, L. *et al.* Complete genome sequence and comparative genomic analyses of the vancomycin-producing Amycolatopsis orientalis. *BMC Genomics* **15,** 363 (2014).

22. Li, T. L. *et al.* Biosynthetic Gene Cluster of the Glycopeptide Antibiotic Teicoplanin: Characterization of Two Glycosyltransferases and the Key Acyltransferase. *Chem. Biol.* **11,** 107–119 (2004).

23. van Wageningen, A. M. A. *et al.* Sequencing and analysis of genes involved in the biosynthesis of a vancomycin group antibiotic. *Chem. Biol.* **5,** 155–162 (1998).

24. Donadio, S., Sosio, M., Stegmann, E., Weber, T. & Wohlleben, W. Comparative analysis and insights into the evolution of gene clusters for glycopeptide antibiotic biosynthesis. *Mol. Genet. Genomics* **274,** 40–50 (2005).

25. Vatlin, A. A., Bekker, O. B., Lysenkova, L. N. & Danilenko, V. N. Draft Genome Sequence of Streptomyces fradiae olg1-1, a Strain Resistant to Nitrone-Oligomycin. *Genome Announc.* **3,** e01252-15 (2015).

26. Baranasic, D. *et al.* Genome Sequences of the Oxytetracycline Production Strain Streptomyces rimosus R6-500 and Two Mutants with Chromosomal Rearrangements. *Genome Announc.* **2,** e00517-14 (2014).

27. Kaur, N., Kumar, S., Bala, M., Raghava, G. P. S. & Mayilraj, S. Draft Genome Sequence of Amycolatopsis decaplanina Strain DSM 44594T. *Genome Announc.* **1,** e0013813 (2013).

28. Owen, J. G. *et al.* Mapping gene clusters within arrayed metagenomic libraries to expand the structural diversity of biomedically relevant natural products. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 11797–802 (2013).

29. Banik, J. J. & Brady, S. F. Cloning and characterization of new glycopeptide gene clusters found in an environmental DNA megalibrary. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 17273–17277 (2008).

30. Banik, J. J., Craig, J. W., Calle, P. Y. & Brady, S. F. Tailoring enzyme-rich environmental DNA clones: A source of enzymes for generating libraries of unnatural natural products. *J. Am. Chem. Soc.* **132,** 15661–15670 (2010).

31. Thaker, M. N. *et al.* Identifying producers of antibacterial compounds by screening for antibiotic resistance. *Nat. Biotechnol.* **31,** 922–927 (2013).

32. Baltz, R. H. Molecular engineering approaches to peptide, polyketide and other antibiotics. *Nat. Biotechnol.* **24,** 1533–1540 (2006).

33. Truman, A. W. *et al.* Antibiotic resistance mechanisms inform discovery: identification and characterization of a novel amycolatopsis strain producing ristocetin. *Antimicrob. Agents Chemother.* **58,** 5687–5695 (2014).

Supplemental Information

Table 1: Glycopeptide and Novel Gene Clusters

No.	Gene Cluster	Acc. Number	Organism	Size (kb)	Database
1	Vancomycin	BGC0000455	Amycolatopsis orientalis HCCB10007	65	MIBiG
2	Balhimycin	BGC0000311	Amycolatopsis DSM 5908	66	
3	A40926	BGC0000289	Nonomuraea sp. ATCC 39727	89	
4	Ristocetin	BGC0000418	Amycolatopsis lurida NRRL2430	80	
5	Ristomycin A	BGC0000419	Amycolatopsis japonica MG417-CF17	68	
6	Teicoplanin	BGC0000440	Actinoplanes teichomyceticus ATCC31131	89	
7	A47934	BGC0000290	Streptomyces toyocaensis NRRL 15009	69	
8	Complestatin	BGC0000326	Streptomyces lavendulae	55	
9	Chloroeremomycin	BGC0000322	Amycolatopsis orientalis	34	
10	Pekiskomycin	JX026280	Streptomyces sp. WAC1420	82	
11	UK-68,597	BGC0001178	Actinoplanes sp. ATCC 53533	80	
12	Predicted GBGC	ANSJ01000022	Streptomyces rimosus subsp. Rimosus ATCC 10970	52	NCBI
13		LGSP01000035	Streptomyces fradiae olg1-1	67	
14		JNAD01000004	Streptomyces fradiae ATCC 19609	82	
15		JJNO01000005	Streptomyces rimosus R6-500	75	
16		JMGX01000001	Streptomyces rimosus R6-500MV9	75	
17		LOBU02000007	Amycolatopsis regifaucium GY080	94	

18		AOHO01000074	Amycolatopsis decaplanina DSM 44594	94	
19		KF264565	uncultured bacterium esnapd26	86	
20		KC688274	Streptomyces sp. WAC4229	69	
21		MQUP01000022	Amycolatopsis keratiniphila subsp. nogabecina FH 1893	131	
22		MUXN01000025	Amycolatopsis azurea DSM 43854	102	
23		MQUQ01000006	Amycolatopsis coloradensis DSM 44225	129	
24		ANMG01000033	Amycolatopsis azurea DSM 43854	67	
25		EU874253	Uncultured soil bacterium clone D30 TEG	51	
26		HM486075	Uncultured organism CA878 CA878	89	
27		LIWC01000005	Amycolatopsis sp. CB00013 CB00013	91	

Table 2: Best Models prediction with Prottest 2.4 Server

```
*************************************************************
Best model according to AIC: WAG+I+G+F
*************************************************************
Model          deltaAIC*    AIC          AICw       -lnL
-------------------------------------------------------------
WAG+I+G+F      0.00         85875.89     1.00       -42761.95
JTT+I+G+F      250.44       86126.33     0.00       -42887.16
LG+I+G+F       276.47       86152.36     0.00       -42900.18
CpREV+I+G+F    276.93       86152.82     0.00       -42900.41
LG+G+F         278.25       86154.14     0.00       -42902.07
```

Figure 1: Workflow for the Identification of Novel GBGCs

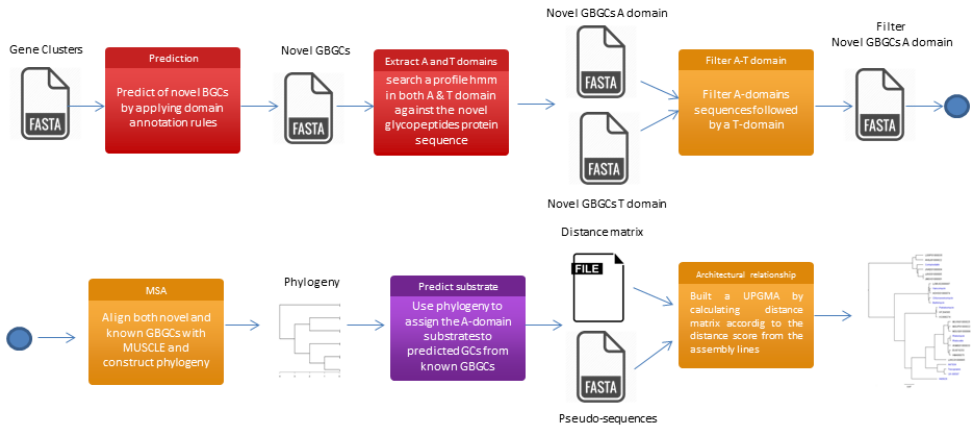

Figure 2: Workflow project chart