Ranking microbial metabolomic and genomic links using correlation-based and feature-based scoring functions

Grímur Hjörleifsson Eldjárn [1], Andrew Ramsay [1], Justin J. J. van der Hooft [2], Katherine R. Duncan [3], Sylvia Soldatou [4], Juho Rousu [5], and Simon Rogers [1,*]

[1] School of Computing Science, University of Glasgow, Glasgow, G12 8RZ, UK,
[2] Bioinformatics Group, Wageningen University, 6708 PB, Wageningen, The Netherlands,
[3] Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, Glasgow, G4 0RE, UK,
[4] School of Pharmacy and Life Sciences, Robert Gordon University, Aberdeen, AB10 7GJ, UK and
[5] Department of Computer Science, Aalto University, Espoo, 02150, Finland.

* To whom correspondence should be addressed.

Abstract

Motivation: Specialised metabolites from microbial sources are well-known for their wide range of biomedical applications, particularly as antibiotics. When mining paired genomic and metabolomic data sets for novel specialised metabolites, establishing links between Biosynthetic Gene Clusters (BGCs) and metabolites represents a promising way of finding such novel chemistry. However, due to the lack of detailed biosynthetic knowledge for the majority of predicted BGCs, and the large number of possible combinations, this is not a simple task. This problem is becoming ever more pressing with the increased availability of paired omics data sets. Current tools are not effective at identifying valid links automatically, and manual verification is a considerable bottleneck in natural product research.

Results: We demonstrate that using multiple link-scoring functions together makes it easier to prioritise true links relative to others. Based on standardising a commonly used score, we introduce a new, more effective, score, and introduce a novel score using an Input-Output Kernel Regression approach. Finally, we present NPLinker, a software framework to link genomic and metabolomic data. Results are verified using publicly available data sets that include validated links.

Availability: NPLinker is available at https://github.com/sdrogers/nplinker/, current version tagged *scoring_jun2020*

Contact: simon.rogers@glasgow.ac.uk

Supplementary information: Supplementary data are available on *bioRxiv*.

1 Introduction

Microbial specialised metabolism, i.e. microbial production of metabolites not strictly needed for the survival of the organism, has been a rich source of metabolites for a variety of biomedical applications (Newman and Cragg, 2020). Recent advances in genomic analysis (Blin *et al.*, 2019; Hannigan *et al.*, 2019; Cimermancic *et al.*, 2014) indicate that microorganisms harbour considerable untapped metabolomic potential (Baltz, 2017).

Genes responsible for the production of microbial specialised metabolites are usually grouped into Biosynthetic Gene Clusters (BGCs), contiguous regions of adjacent genes that, taken together, encode enzymes that produce one or several structurally related metabolites. It is increasingly common to simultaneously explore the biosynthetic potential of multiple related strains. When doing so, it can be informative to group BGCs into Gene Cluster Families (GCFs). This clustering is based on defining a similarity-based distance between BGCs (Goering *et al.*, 2016; Navarro-Muñoz *et al.*, 2020) with the assumption that BGCs that are close with regards to this distance will produce identical or structurally similar metabolites. The current state-of-the-art tools for microbial BGC detection and subsequent clustering are antiSMASH (Blin *et al.*, 2019) and BiG-SCAPE (Navarro-Muñoz *et al.*, 2020), respectively. Many predicted BGCs encode for unknown products, and even when the products are known, not all BGCs are expressed under normal laboratory conditions (Baltz, 2017).

Microbial specialised metabolites that are expressed under laboratory conditions are usually profiled using tandem mass spectrometry (MS/MS or MS2). Using this technique, each ionizable metabolite in a culture extract has a corresponding spectrum, representing the fragmentation of the metabolite. Data from multiple strains can be combined allowing researchers to see in which strains particular spectra appear. Structurally similar metabolites are assumed to give similar fragmentation patterns, and this can be exploited by clustering similar spectra into Molecular Families (MFs) in a process called *molecular networking* (Watrous *et al.*, 2012).

Linking microbial specialised metabolites to their producing BGCs has until recently mostly been done manually and on a small scale, working with single strains, either based on similarity to known, existing links (e.g. Xu *et al.* (2018)), or predictions of unique identifying features of the spectrum from the BGCs (e.g. Kaweewan *et al.* (2017)). However, to increase the chance of successfully finding links for useful or

© The Author 2020. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

interesting metabolites, this has to be done at a large scale. The increased throughput makes manual methods prohibitively labour intensive, making development of computational methods for this purpose necessary.

Existing computational approaches to this problem can be broadly classified into two categories: *feature-based approaches*, where the set of MS2 spectra is searched for predicted structural features of the putative product of the BGC (*peptidogenomcs*, Kersten *et al.* (2011)), and *correlation-based approaches*, where similarities in sets of strains containing specific BGCs on the one hand, and specific spectra on the other hand, are used to evaluate the links between BGCs and spectra (*metabologenomics*, Goering *et al.* (2016); Doroghazi *et al.* (2014)).

Currently, various tools exist to aid in *feature-based linking*, such as SANDPUMA (Chevrette *et al.*, 2017), GNP (Johnston *et al.*, 2015) and RippQUEST (Mohimani *et al.*, 2014). These tools are each designed to be used for only a single product type such as NRPS, PKS or RiPP, respectively. Further analysis is therefore required to evaluate the links from different tools against each other. *Correlation-based* scores suffer from a distinct size-based discrepancy between links involving GCFs and MFs of different sizes, as will be discussed in Section 2.1. Furthermore, little research has been done into how the various scoring methods correlate or complement one another.

For detailed reviews of linking microbial specialised metabolites to their producing BGCs, the reader is referred to Soldatou *et al.* and van der Hooft *et al.* (Soldatou *et al.*, 2019; van der Hooft *et al.*, 2020).

Central to metabologenomics is the clustering of BGCs into GCFs and MS2 spectra into MFs. The question of which spectrum is linked to which BGC can then be reframed as either which GCF is linked to which MF or which GCF is linked to which spectrum (both of which have been successfully used for individual links, e.g. Duncan *et al.* (2015) and Goering *et al.* (2016), respectively).

Identifying the correspondence between BGCs and MS2 spectra can be useful in several ways. Since rediscovery of known metabolites is a persistent problem in metabolomics (Mohimani *et al.*, 2017), linking a spectrum to a BGC with a known metabolite product can help identify the spectrum as belonging to that metabolite, and thus complement database matching as a dereplication strategy. An example of this approach can be seen in the characterisation of a polyketide antibiotic with an elemental formula of $C_{35}H_{56}O_{13}$ (Mohimani *et al.*, 2018). Secondly, since looking for metabolites similar to a known metabolite is a common starting point in microbial metabolomics (e.g. Duncan *et al.* (2015)), establishing links between a BGC similar to a BGC with a known metabolite product, and a spectrum, can be considered particularly useful. This strategy, for example, was used to isolate the novel lanthipeptide tikitericin from a thermophilic bacterium (Xu *et al.*, 2018). Lastly, since the expression of the metabolite is sometimes controlled by genes within the BGC, e.g. transporter- and regulatory genes, knowing which BGC produces a particular metabolite can give valuable insight for synthetic biology applications, e.g. heterologous expression. This approach was used to identify the normally cryptic BGC responsible for the production of scleric acid, a secondary metabolite with moderate antimicrobial and antitumor activity (Alberti *et al.*, 2019).

For a collection of BGCs in a strain, and a collection of metabolites produced by the same strain, a priori, any given metabolite could potentially be produced by any of the BGCs. One approach to figuring out which BGC is in fact producing which metabolite is to rank these potential links according to how likely each BGC is to produce each metabolite. This is equivalent to assigning each hypothetical BGC-metabolite link a score, and then ordering the list of links according to the score. Ideally, the resulting ranking would put all true BGC-metabolite links above others.

In this paper, we consider the problem of ranking potential links between BGCs and MFs (or MS2 spectra, which can be considered as singleton MFs) in a data set consisting of paired genomic and metabolomic data for multiple, similar strains, to prioritise them for further investigation.

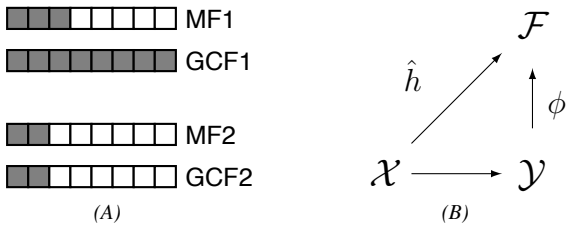

Fig. 1. (A) Size discrepancy in the strain correlation score for GCFs of varying sizes. Each box represents a strain, with filled boxes denoting that the strain is a member of the GCF or MF, and blank boxes that it is not. The top GCF-MF pair outscores the bottom pair by 30 to 26, despite the top pair having arguably stronger correspondence. (B) Arrow diagram of the IOKR framework. $\mathcal{X}$ denotes the space of MS2 spectra, $\mathcal{Y}$ is the space of metabolites, and $\mathcal{F}$ is the shared space of molecular fingerprints. $\hat{h}$ is the (learned) mapping from MS2 to fingerprints, while ϕ is the (exact) mapping from metabolites to molecular fingerprints.

Our contribution to the field is as follows: we demonstrate that by using multiple, complementary scoring functions we can more effectively prioritise valid links between GCFs and MFs, than using individual scoring functions. To this end, we start by investigating the behaviour of the strain correlation score introduced in (Doroghazi *et al.*, 2014), especially with regards to different sizes of GCFs and MFs. We demonstrate that standardising this score makes it more effective, assigning previously validated links higher scores relative to all other potential links. We also show how Input-Output Kernel Regression (Brouard *et al.*, 2016a), which has originally been developed for metabolite annotation, can be used as a new feature-based method for ranking BGC-MF links. Finally, we present NPLinker, a software framework to link genomic and metabolomic data within which the link-scoring methods introduced here have been implemented. Our methods are tested on published paired omics data sets including validated BGC-MS2 links, and we show an improved ability to prioritise the previously validated links relative to other potential links in the data set.

2 Methods

2.1 Strain correlation scoring

Consider a population of strains, each with a set of predicted BGCs. The union of those sets constitutes a set of BGCs in the population. Assume also that the BGCs have been clustered into GCFs. Similarly, assume that associated with the population is a set of all MS2 spectra for metabolites produced by the population, clustered into MFs. We are interested in scoring these potential GCF-MF links, in order of how likely each GCF is to produce the metabolite giving rise to each MF.

The most common scoring method for scoring a GCF-MF link in metabologenomics is as follows (Doroghazi *et al.*, 2014): starting from zero, for each strain in the population,

- add 10 to the score if the strain produces the metabolite and has a BGC in the GCF,
- subtract 10 from the score if the strain produces the metabolite but does not have a BGC in the GCF,
- add 1 to the score if the strain neither produces the metabolite nor has a BGC in the GCF, and
- leave the score unchanged if the strain has a BGC in the GCF but does not produce the metabolite.

Considering the GCF as a set G of strains that contribute a BGC to the GCF, and the MF as a set M of strains that produce a metabolite in the MF, the score is therefore defined as a function $\sigma(M, G)$ which depends on the size of M ($\#M$), the size of G ($\#G$), and the overlap between the two sets, in addition to a set of weights (here 10, -10, 1 and 0, respectively).

While intuitive and easy to interpret, this scoring function, hereafter referred to as *strain correlation score*, has the major drawback of being highly dependent on both total population size and, with potentially greater impact, the size of the GCF and the number of strains that produce the metabolite. This makes comparison between possible links involving GCFs and MFs of different sizes challenging, even within the same data set.

2.2 Standardisation of the strain correlation score

Using the strain correlation score, different sizes of MFs and GCFs greatly influence the link score. A link between a GCF that contains a BGC from every strain in the data set and a moderately-sized MF can easily outscore a link between a smaller GCF and MF that have perfect strain correspondence, even if the evidence for the latter link being valid is stronger, as shown in Figure 1.

This effect can be mitigated by standardising the score. For a given GCF and MF, let G and M be the sets of strains contributing to the GCF and the MF respectively. Assuming the null hypothesis that the strain content of G and M is not correlated, the probability of a given overlap between the two follows a hypergeometric distribution with the total number of strains as the population size N, the size of the molecular family M as the number of positives in the population, the size of the GCF G as the sample size, and the size of the overlap between the two as the number of positives in the sample.

Letting $m = \#M$, $g = \#G$, $o = \#(G \cap M)$ and $n = \#N$, we can define the strain correlation $\sigma(M, G) = \sigma'(m, g, o, n)$ in terms of the sizes of M, G and the overlap between the two. We can therefore compute the expected value $E[c(M, G)]$ as a function of the sizes of M and G as

$$E[\sigma(M, G)] = \sum_k p(o = k)\sigma'(m, g, k, n)$$

where k runs over all possible sizes of $\#(M \cap G)$, and $p(o = k)$, the probability of the size of the overlap $\#(M \cap G)$ being k, follows the hypergeometric distribution as previously stated. The variance can then be computed as

$$Var[\sigma(M, G)] = E[\sigma(M, G)^2] - E[\sigma(M, G)]^2$$

Figure 2 shows the expected value and variance of the strain correlation score as a function of the sizes of G and M, for population size $\#N = 100$. The expected value varies greatly, especially with the number of strains producing the metabolite (due to the higher magnitude coefficients when the spectrum is present in the strain than when it is absent), and the variation in variance across the domain is also considerable.

We propose to mitigate this by defining a *standardised strain correlation score* for a GCF G and MF M as

$$\bar{\sigma}(M, G) = \frac{\sigma(M, G) - E[\sigma(M, G)]}{Var[\sigma(M, G)]}$$

In this version, the score for each prospective link takes into account the sizes of the GCF and the MF and adjusts the score accordingly, making the scores comparable between links involving strain sets of different sizes as is necessary when, for example, comparing scores for different spectra / MF for a particular GCF. In the case of Figure 1, the standardised scores are 0.0 and 2.65, favoring the right-hand pair.

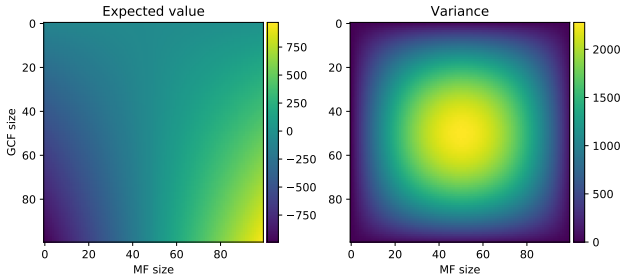

Fig. 2. Expected value and variance of the strain correlation score for a population of 100 strains, as a function of GCF and MF sizes. Both the expected value and the variance have a considerable range, rendering comparison between links involving different sizes of GCFs and MFs difficult.

2.3 Input-Output Kernel Regression

If a BGC is known to produces a certain metabolite, the problem of linking a spectrum to that BGC is equivalent to linking a spectrum to the metabolite that the BGC produces. This problem, of matching spectra to molecular structures, is an important problem in metabolomics because it underpins all untargeted metabolomics workflows. This is often done in the space of *molecular fingerprints*, which are binary vectors denoting the various properties of the molecule, including the presence or absence of certain substructures. These fingerprints can be derived from candidate structures and effectively predicted from spectra.

Brouard *et al.* propose Input-Output Kernel Regression (IOKR) (Brouard *et al.*, 2016a,b) as a method of ranking a candidate set of chemical structures, given an input spectrum. They compare IOKR with state of the art methods such as CSI:FingerId (Dührkop *et al.*, 2015) and demonstrate similar performance with considerably shorter training and classification times, with the training time reduced from 82 hours to under one minute, and classification time reduced by half.

In principle, IOKR works by mapping similarities in the input space (spectra) and the output space (molecular structures) to the molecular fingerprint space and searching a candidate set of metabolites for the closest match in that space. Let $\mathcal{X}$ be the space of MS2 spectra and $\mathcal{Y}$ the space of metabolites, with kernel functions $\mathcal{K}_x, \mathcal{K}_y$, respectively, and $X \subset \mathcal{X}$ and $Y \subset \mathcal{Y}$ be training sets of paired metabolites and spectra, where each $x_i \in X$ has a corresponding element $y_i \in Y$, and vice versa. Let $\mathcal{F}$ be the space of molecular fingerprints, and $\hat{h} : \mathcal{X} \to \mathcal{F}, x \mapsto \sum_{x_j \in X} \mathcal{K}_x(x, x_j)\mathbf{c}_j$, where $\mathbf{c}_j \in \mathcal{F}$. Finally, let $\phi : \mathcal{Y} \to \mathcal{F}$ be the mapping expressing the kernel function $\mathcal{K}_y$ as the inner product in $\mathcal{F}$, $\mathcal{K}_y(\cdot, \cdot) = \langle \phi(\cdot), \phi(\cdot) \rangle_{\mathcal{F}}$. Given a spectrum x, IOKR works by searching a set Y^* of candidate structures for an element y that minimises the expression $\langle \hat{h}(x), \phi(y) \rangle_{\mathcal{F}}$.

For further details of the underlying mathematics of IOKR, please refer to Brouard *et al.* (2016b).

2.4 Using IOKR to rank BGC-spectrum links

In the context of linking BGCs and spectra, IOKR can be considered a *feature-based method*. When scoring links between individual BGCs and spectra, however, it does not predict the features of the spectrum directly from the BGC, but uses molecular fingerprints as an intermediate.

The application of IOKR depends heavily on the choice of kernel function on the spectra, and on the choice of molecular fingerprint. For this work, we choose kernel functions that are easy and fast to compute, with the caveat that further optimisation may be possible.

Molecular fingerprints are extracted from SMILES strings using the Chemistry Development Kit (Willighagen *et al.*, 2017). The fingerprint

vector is composed of three concatenated sets of fingerprints: *CDK Substructure*, *PubChem Substructure* and *Klekota-Roth* fingerprints. Taken together, these cover most of the molecular properties covered by the fingerprint used by Brourad *et al.* and result in similar performance (Brouard *et al.*, 2016a).

As a denoising step, to avoid time-consuming computation of fragmentation trees for the spectra, we filter the input spectra to include only the peaks found in the training data, before using the *Probability Product Kernel* (PPK) (Jebara *et al.*, 2004; Brouard *et al.*, 2016a).

Since the standardised strain correlation score is defined between MFs and GCFs, and the IOKR score between BGCs and spectra, they are not directly comparable. To be able to use them together, we generalise the IOKR score to GCF-MF links by taking the highest BGC-spectrum pair where the BGC is in the GCF and the spectrum in the MF, and assign that score to the GCF-MF pair, i.e. for a GCF G and a MF M, and a scoring function σ scoring BGC-spectrum links, we define a second generalised function $\hat{\sigma}$ by setting $\hat{\sigma}(M, G) = max_{m \in M, g \in G} \sigma(m, g)$.

2.5 Combining strain correlation and feature-based scores

So far, computational efforts to find GCF-MF links have mostly been based on either the *feature-based* or the *correlation-based* approach. However, both approaches have limitations. For example, correlation-based methods have trouble prioritising singleton links within the same strain, i.e. when the same strain is associated with multiple singleton GCFs and MFs. Feature-based methods on the other hand rely on being able to predict distinguishing features that can be detected in MS2 spectra from the BGCs, and since the same features are often present in multiple GCFs and MFs, this can yield multiple potential links with the same score.

Because the two approaches are based on very different principles, they are likely to be at least in part complementary, with each one sensitive to things that the other is not. We demonstrate that combining the two approaches by looking at potential links that are highly ranked using both increases the ratio of previously validated links to all links in the joint top percentiles compared to using either one of the approaches.

2.6 NPLinker

To facilitate analysis of paired genomics and metabolomics data sets, we developed NPLinker, a Python module to accelerate and support the process of automatically linking GCFs or their BGCs with observed mass spectra. NPLinker accepts genomic outputs from antiSMASH and BiG-SCAPE (including MIBiG reference BGCs), and metabolomic output from GNPS. Additionally, it includes integration with the *Paired omics Data Platform* to retrieve paired public genomics and metabolomics data (https://pairedomicsdata.bioinformatics.nl).

NPLinker creates objects for spectra, MFs, BGCs and GCFs in the data set, maintaining the hierarchical relationship between them, and keeps track of strain ID or IDs associated with each object, as well as strain aliases. Various scoring functions can be used to evaluate the links between metabolomic and genomic objects, both scoring functions that are supplied with NPLinker and custom scoring functions. Objects can also be filtered by various criteria such as strain association, inclusion in MIBiG, and annotations.

2.7 Data sets

2.7.1 Training data

The IOKR model is trained on a set of spectrum-molecular fingerprint pairs. Since molecular fingerprints can be computed from structural annotations, such as SMILES (Weininger, 1988) or InChI-strings (Heller *et al.*, 2015), we use library MS2 spectra from the public, community-driven GNPS knowledge base (Wang *et al.*, 2016) as a training set for the IOKR model. We use the same training data set as Brouard *et al.* (Brouard *et al.*, 2016a), which consists of 4138 spectra from GNPS with structural annotations. The spectra are for metabolites from a variety of sources, including microbial, plant and human metabolites.

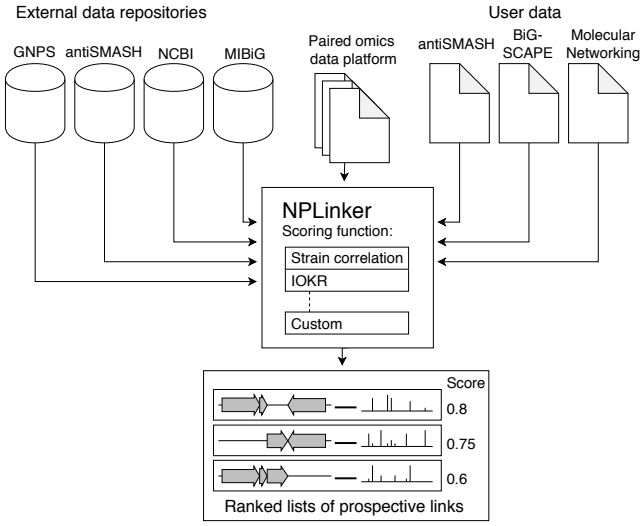

Fig. 3. The NPLinker module helps with automatically linking GCFs and MFs. It integrates metabolomic and genomic data sets, using either external sources, user-provided data, or a mixture of both, and ranks potential links between metabolomic and genomic objects by given scoring functions, either built-in or user-defined.

2.7.2 Test data

In recent years, the MIBiG database (Medema *et al.*, 2015) has emerged as a central repository of characterised microbial BGCs. The current version, 2.0 (Kautsar *et al.*, 2019), contains close to 2000 BGCs, most of which have structural annotations. Many tools, including antiSMASH and BiG-SCAPE, use MIBiG as part of their analysis to quantify similarity of unknown to known BGCs.

In particular, antiSMASH can be configured to compute the similarity of detected BGCs to MIBiG entries and return for each detected BGC a (ranked) list of similar BGCs in MIBiG, using the *known cluster blast* feature (Blin *et al.*, 2019). Assuming that similar BGCs give rise to similar compounds, we used this list in turn to assign one or more molecular structures to BGCs, according to how many high-scoring matches are found in MIBiG (or none, if no match is found).

In its current form, MIBiG has no information about the MS2 spectra of the metabolites produced by the BGCs. However, the entries are annotated with structures, so the structural annotations included in GNPS can be used to link GNPS spectra to their corresponding MIBiG BGCs. In this way, we built a set of known BGC-spectrum pairs. To avoid distinguishing between metabolites based on properties absent from an MS2 spectrum, e.g. the chirality of the metabolite, this linking is done using only the first part of the InChI-key of each metabolite. This yields 2966 BGC-spectrum pairs, each with an associated metabolite, which can be used to evaluate the IOKR model proposed in this paper. These pairs include 2069 unique spectra and 242 unique MIBiG BGCs. This list is included in the Supplimentary Information.

A major problem in the development of methods to link BGCs to spectra for collections of microbial strains is the lack of ground truth data. In an attempt to ameliorate the situation, Schorn *et al.* recently developed the Paired omics Data Platform documenting the location of genomic and metabolomic data sets from microbial experiments, with a focus on data sets with BGCs and MS2 spectra (Schorn *et al.*, 2020). This

gives a repository of established links in various data sets which can then be used to evaluate scoring methods for prospective links between BGCs and spectra, in terms of relative over-representation of established links towards the upper end of the distribution of scores. From this platform we concentrated on three data sets: MSV000078836 (Crüsemann *et al.*, 2017), MSV000085018 (Gross *et al.*, 2007) and MSV000085038 (Leao *et al.*, 2017), hereafter referred to as Crüsemann, Gross and Leao, respectively.

The Crüsemann data set consists of 120 microbial strains with 8 established links between a BGC and a MF, the Gross data set consists of 7 strains with 9 established links between a BGC and a MF, and the Leao data set contains 4 strains with 5 established links between a BGC and a specific MS2 spectrum. After downloading the strain assemblies and metabolomic data, the genomes were run through antiSMASH v5.0.0 for BGC detection and BiG-SCAPE v1.0.0 to cluster the BGCs into GCFs.

The paired data platform links are represented by molecular families or MS2 spectra in specific GNPS data sets, and by MIBiG IDs in specific strains. To map the BGC links back to the genomes in question we used antiSMASH to score the correspondence between the MIBiG entries and the detected BGCs, returning for each BGC a (possibly empty) list of MIBiG matches, along with a score for each match.

3 Results

3.1 Standardising the strain correlation score

As the Crüsemann data set contained the largest number of strains of the three data sets being considered, it was selected to evaluate the standardised strain correlation score. To do this, we examine the distribution of scores for validated links in relation to the scores for all hypothetical links. The more effective a scoring method is, the higher it should rank the correct BGC-MS2 links compared to the rest of the links. Therefore, the mean score for validated links should be higher than the mean score for all potential links, so the null hypothesis for testing the validity of the scoring function is that both distributions of scores (for validated links and all links) have the same mean.

For the raw strain correlation score, the mean score is 83.514 for all links, and 14.667 for validated links, i.e. a lower mean score for the validated links than for all links, which is the opposite of what would be expected. Standardising the score gives a mean score of -0.006 for all links, and 3.672 for validated links (Table 1). The distributions of the scores of the validated links amongst all links are shown in Figure 4 for both the standardised and raw versions of the score. Similar trends can be observed in the Leao and Gross data sets, please see Supplementary Information. The results demonstrate the importance of our proposed standardisation process.

Columns three and four of Table 3 show the proportion of verified links among the top scoring links for the raw and standardised correlation scores. In the Crusemann data set, where the strain correlation scoring is most relevant, the proportion is considerably higher for the standardised correlation score, relative to the raw correlation score.

3.2 Evaluating the IOKR scoring function using MIBiG

As stated in Section 2.7.1, the training set used to build the IOKR model includes metabolites from sources other than microbial. As the performance evaluation of IOKR does not break down performance by metabolite sources, to evaluate the performance of IOKR on microbial specialised metabolites specifically, we tested the method on the paired MIBiG/GNPS data (see Section 2.7.2) by matching each spectrum to the candidate set consisting of all structures associated with an MIBiG entry. As multiple BGCs can produce the same metabolite, and as the same BGC can be associated with a number of structurally related metabolites, this is a many-to-many relation.

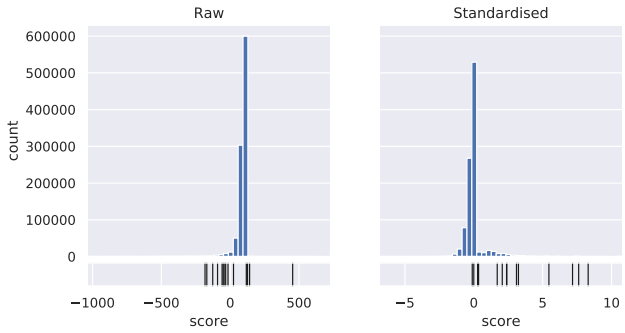

Fig. 4. Distribution of the raw and standardised strain correlation scores, as well as the distribution of the scores for validated links (in black) relative to the distribution of scores for all links, in the Crüsemann data set. The standardised score has a more pronounced tail at the top end, which includes 13 out of 15 validated links, whereas many of the validated links score relatively low on the distribution of the raw scores.

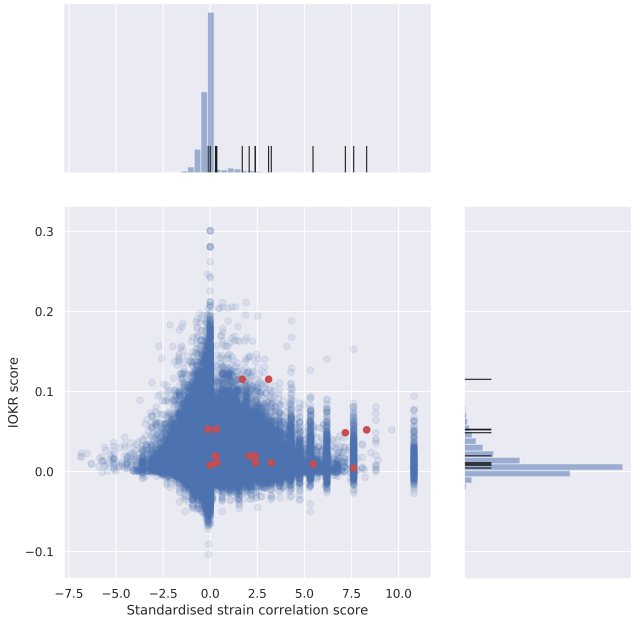

Fig. 5. IOKR- and strain correlation scores for all potential links in the Crüsemann data set, with histograms of the scores. Verified links are coloured green on the joint plot, and marked on the distribution histograms. Verified links are concentrated in the upper-right quadrant, i.e. score relatively high on both axes. See Supplementary Information for results for the two further data sets.

For each spectrum, IOKR returns an ordered list of all metabolites in the candidate set. As multiple MIBiG BGCs can be associated with the same metabolite, the correct metabolite for the spectrum gives rise to one or more spectrum-BGC links.

To translate the ranking of the metabolites into the rank of a BGC among a set of BGCs, we consider the highest-ranking metabolite that the BGC produces. Each metabolite that scores higher than the correct metabolite is associated with one or more BGC. The rank of the correct BGC is the number of distinct BGCs associated with a metabolite that scores higher than the correct metabolite.

For comparison purposes, a baseline score was estimated by randomising the rank of the structures for each spectrum, and the same process was repeated to assign a BGC using the randomised score.

Table 1. The mean score for all links and the subset of validated links in the Crüsemann data set.

Method	all	validated	p-value
Raw correlation	83.5144	14.6667	0.0001
Standardised correlation	-0.0060	3.6717	6.8302e-64
IOKR	0.0105	0.0364	1.7968e-9

Table 2. Top-n accuracy, and AUC, of IOKR on MIBiG data.

	top-1	top-5	top-10	top-20	top-200	AUC
data	0.1208	0.1708	0.1870	0.2121	0.2946	0.6534
random	0.0	0.0014	0.0044	0.0103	0.1486	0.5209

Table 3. Proportion of verified links among all possible GCF-MF links in the three data sets.

Data set	Total	Top raw corr.	Top std. corr.	Top IOKR	Top comb
Crüsemann	1.50096e-5	4.43106e-06	0.00021	0.00012	**0.00156**
Gross	9.96242e-6	3.48371e-05	7.69024e-5	1.99247e-5	**0.00019**
Leao	0.00086	0.00466	0.00385	0.00211	**0.01298**

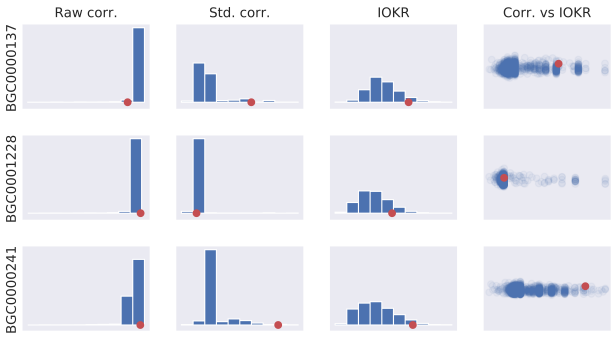

Fig. 6. Position of the score for the validated GCF-MF pair (red dot) within the distribution of the scores of the links between that particular GCF and all MFs, for a selection of established links in the Crüsemann data set (rows). The first three columns show histograms of the raw and standardised versions of the strain correlation score, as well as the IOKR score, for all links including a given GCF, with the score of the correct link indicated. The last column shows the standardised correlation score (x-axis) and IOKR score (y-axis) for the same links, again with the correct link indicated. Both IOKR and the standardised correlation scores tend to put established links higher in the distribution of scores for the GCF in consideration, than the raw correlation score. Furthermore, some of the validated links score relatively higher on IOKR than the standardised strain correlation score, and vice versa, suggesting that the two scores complement one another. For full results, as well as for other data sets, please refer to Supplementary Information.

Table 2 shows the top-n performance of IOKR, i.e. how often the 'true' BGC match for a given spectrum is among the top n matches returned by IOKR, for a selection of n. IOKR outperforms the baseline by a considerable margin, especially at low values of n, with an AUC (i.e. the area under the top-n curve for varying values of n) of 0.6534 compared to 0.5209 for the null distribution. In this context, the AUC can be considered as a measure of how close we are to the top hit being the correct one in all instances (AUC of 1.0) compared to a randomised baseline.

3.3 Evaluating the performance of IOKR

Similarly to the evaluation of the standardised strain correlation score, we can observe the distribution of the scores for the validated links among the scores for all potential links in the Crüsemann data set. Out of 3316 BGCs in the data set, 2242 could be assigned structure based on similarity to MIBiG entries, and used as candidate set for the 6246 MS2 spectra in the data set. As can be seen from Figure 5, the upper end of the distribution for the IOKR score contains a relatively high proportion of the validated links, with the mean score of 0.0105 for all links and 0.0364 for validated links (Table 3). Results for other data sets can be found in Supplementary Information.

3.4 Complementarity of IOKR and strain correlation scoring

As the aim of scoring the potential links between a BGC and a spectrum in a given data set is to prioritise links for further verification, using multiple scoring functions to complement one another can be helpful in the prioritisation, as not all high-scoring links are likely to be valid.

To demonstrate the efficacy of the standardised strain correlation and IOKR scores, we can look at the proportion of verified links to potential links scoring above a given percentile for either or both of the scores, and observe that the higher ends of the distributions are enriched for verified links, i.e. contain a relatively higher number of verified links compared to the whole data set. Furthermore, to demonstrate that they are in fact complementary, we can observe that the set of links that have both high IOKR and strain correlation scores contains an even higher relative number of verified links than the individual scores. Table 3 shows the proportions of verified links for the three data sets investigated. The first column contains the ratio of verified links to all possible links, the second and third contain the same ratio for links with IOKR and standardised strain correlation scores above the 95th percentile respectively, and the last column contains the same ratio for links with both strain correlation and IOKR score above 95th percentile of the respective scores. As no link in the Gross data set scored above 95th percentile on both IOKR and strain correlation scores, the numbers in the table for that data set are for the 90th percentile. The proportion of verified links among high-scoring links for each score is higher by an order of magnitude compared to all links, and higher again by an order of magnitude if both scores are used for filtering. The distribution of the scores for the Crüsemann data set, and the relative score of the verified links, can be seen in the right-hand histogram on Figure 5.

Since the number of verified links in each data set is small (ranging from 5 to 15), we can pool the scores to get a clearer sense of the statistical significance. Considering the 95th percentile per data set for both scores, and adding up the numbers of links in each category (verified or unverified, and scoring above 95th percentile for either or both scores) for all the data sets, both the IOKR and strain correlation scores are significantly enriched (p-value of 2.409×10^{-4} and 1.164×10^{-11}, respectively) for validated links compared to all links scoring above their respective 95th percentiles. Furthermore, the set of links scoring above the 95th percentile on both is significantly enriched compared to either of the individual scores (p-value of 3.404×10^{-4} and 3.233×10^{-3}, respectively).

3.5 Scoring potential links for a particular BGC

A common approach to establishing correspondence between a GCF and a metabolite is to start with a GCF with established homology to a particular MIBiG BGC. The question then becomes one of ranking the potential GCF-MF links *for that particular GCF* to find the most likely product. Figure 6 shows the position of a selection of validated GCF-MF links in the Crüsemann data set relative to the distribution of all scores for that particular GCF. Similar results for other links, as well as other data sets, can be found in Supplementary Information.

As can be seen, the standardised correlation score does a far better job of prioritising the correct link than the raw score, and is nicely complemented by the IOKR score in some of the cases where both strain correlation scores fail to assign the correct link a relatively high score.

4 Discussion

We have shown that standardising the strain correlation score makes it more effective at prioritising validated links relative to all links, and introduced IOKR as a complementary feature-based scoring function. The major strength of IOKR combined to other feature-based methods is that it is not dependent on product type, but yields scores that are directly comparable between different product types. We have also demonstrated that considering both scoring methods at the same time increases the ratio of validated links among high-scoring links. Furthermore, by pairing the MIBiG and GNPS databases, and using the Paired omics Data Platform, we introduced data sets to test the efficiency of the scoring methods, both separately and combined.

The standardised strain correlation score still suffers from the drawback inherent in correlation-based scoring, of not being able to distinguish between potential links showing the same pattern of strain presence or absence. An obvious example is prioritising multiple singleton GCFs and MFs for the same strain. As a complementary scoring function to the strain correlation score, IOKR does not have this limitation. In contrast to other feature-based scoring methods, it also has the benefit of not being dependent on product type, although due to insufficient test set size, breaking down performance by product type is currently difficult. A drawback of the current IOKR scoring method is its reliance on MIBiG homology to assign molecular structures to BGCs, which is needed to compute the molecular fingerprint. This restricts its use to those BGCs which show considerable homology with MIBiG entries. While still useful in this form, predicting molecular fingerprints directly from BGCs would broaden the applicability of the scoring function.

As stated in Section 2.4, IOKR is also highly dependent on the choice of both kernel function and molecular fingerprints. As the molecular fingerprints denote particular substructures of the molecules, creating additional fingerprints that specifically target molecular substructures seen in secondary metabolites may improve performance. Similarly, the kernels used on the MS2 space can almost surely be further optimised, both through Multiple Kernel Learning (Brouard *et al.*, 2016b) and novel approaches such as vector embeddings.

5 Conclusion

Leveraging the correspondence between data from multiple microbial strains is an important tool to aid in linking BGCs to spectra. Standardising the strain correlation score by taking into account the number of strains involved in the link makes the score more useful by minimising the effect of strain count on the score, thereby no longer favouring common BGCs and molecules. We also introduced IOKR as a novel feature-based scoring method for potential BGC-spectrum links, and showed how the two methods complement one another by demonstrating the relative enrichment of previously validated BGC-spectrum links amongst the top-scoring links in both scoring functions. By using both scores simultaneously, the prioritisation of hypothetical links can be made more effective. Finally, we introduced the NPLinker framework to aid in prioritising BGC-spectrum links for further research. We believe that our work provides the natural products community with new tools that ease the combined analysis of genome and metabolome mining approaches. This may pave the way toward the discovery of novel chemistry that is much needed to fight off pathogenic bacteria, viruses, and insects.

Acknowledgements

JJJvdH is thankful for the support of an ASDI grant from the Netherlands eScience Center - NLeSC (grant no. ASDI.2017.030). AR, KRD and SR acknowledge funding from the Biotechnology and Biological Sciences Research Council (BB/R022054/1). KRD, SR and SS are supported by a Carnegie Trust Collaborative Research Grant. JR acknowledges funding from the Academy of Finland (grants 310107 and 334790) and Scottish Informatics and Computing Science Alliance (SICSA) distinguished visiting fellow scheme.

References

Alberti, F., Leng, D. J., Wilkening, I., Song, L., Tosin, M., and Corre, C. (2019). Triggering the expression of a silent gene cluster from genetically intractable bacteria results in scleric acid discovery. *Chem. Sci.*, **10**(2), 453–463.

Baltz, R. H. (2017). Gifted microbes for genome mining and natural product discovery. *J. Ind. Microbiol. Biotechnol.*, **44**(4-5), 573–588.

Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S. Y., Medema, M. H., and Weber, T. (2019). antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.*, **47**(W1), W81–W87.

Brouard, C., Shen, H., Dührkop, K., d'Alché Buc, F., Böcker, S., and Rousu, J. (2016a). Fast metabolite identification with input output kernel regression. *Bioinformatics*, **32**(12), i28–i36.

Brouard, C., Szafranski, M., D'alché-Buc, F., D', F., Brouard, A.-B., and D'alché-Buc, S. (2016b). Input output kernel regression: Supervised and Semi-Supervised structured output prediction with Operator-Valued kernels. *J. Mach. Learn. Res.*, **17**(1), 1–48.

Chevrette, M. G., Aicheler, F., Kohlbacher, O., Currie, C. R., and Medema, M. H. (2017). SANDPUMA: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across actinobacteria. *Bioinformatics*, **33**(20), 3202–3210.

Cimermancic, P., Medema, M. H., Claesen, J., Kurita, K., Wieland Brown, L. C., Mavrommatis, K., Pati, A., Godfrey, P. A., Koehrsen, M., Clardy, J., Birren, B. W., Takano, E., Sali, A., Linington, R. G., and Fischbach, M. A. (2014). Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell*, **158**(2), 412–421.

Crüsemann, M., O'Neill, E. C., Larson, C. B., Melnik, A. V., Floros, D. J., da Silva, R. R., Jensen, P. R., Dorrestein, P. C., and Moore, B. S. (2017). Prioritizing natural product diversity in a collection of 146 bacterial strains based on growth and extraction protocols. *J. Nat. Prod.*, **80**(3), 588–597.

Doroghazi, J. R., Albright, J. C., Goering, A. W., Ju, K.-S., Haines, R. R., Tchalukov, K. A., Labeda, D. P., Kelleher, N. L., and Metcalf, W. W. (2014). A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat. Chem. Biol.*, **10**(11), 963–968.

Dührkop, K., Shen, H., Meusel, M., Rousu, J., and Böcker, S. (2015). Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl. Acad. Sci. U. S. A.*, **112**(41), 12580–12585.

Duncan, K. R., Crüsemann, M., Lechner, A., Sarkar, A., Li, J., Ziemert, N., Wang, M., Bandeira, N., Moore, B. S., Dorrestein, P. C., and Jensen, P. R. (2015). Molecular networking and pattern-based genome mining improves discovery of biosynthetic gene clusters and their products from salinispora species. *Chem. Biol.*, **22**(4), 460–471.

Goering, A. W., McClure, R. A., Doroghazi, J. R., Albright, J. C., Haverland, N. A., Zhang, Y., Ju, K.-S., Thomson, R. J., Metcalf, W. W., and Kelleher, N. L. (2016). Metabologenomics: Correlation of microbial gene clusters with metabolites drives discovery of a nonribosomal peptide with an unusual amino acid monomer. *ACS Cent Sci*, **2**(2), 99–108.

Gross, H., Stockwell, V. O., Henkels, M. D., Nowak-Thompson, B., Loper, J. E., and Gerwick, W. H. (2007). The genomisotopic approach: a systematic method to isolate products of orphan biosynthetic gene clusters. *Chem. Biol.*, **14**(1), 53–63.

Hannigan, G. D., Prihoda, D., Palicka, A., Soukup, J., Klempir, O., Rampula, L., Durcak, J., Wurst, M., Kotowski, J., Chang, D., Wang, R., Piizzi, G., Temesi, G., Hazuda, D. J., Woelk, C. H., and Bitton, D. A. (2019). A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res.*, **47**(18), e110.

Heller, S. R., McNaught, A., Pletnev, I., Stein, S., and Tchekhovskoi, D. (2015). InChI, the IUPAC international chemical identifier. *J. Cheminform.*, **7**, 23.

Jebara, T., Kondor, R., and Howard, A. (2004). Probability product kernels. *J. Mach. Learn. Res.*, **5**, 819–844.

Johnston, C. W., Skinnider, M. A., Wyatt, M. A., Li, X., Ranieri, M. R. M., Yang, L., Zechel, D. L., Ma, B., and Magarvey, N. A. (2015). An automated Genomes-to-Natural products platform (GNP) for the discovery of modular natural products.

Nat. Commun., **6**, 8421.

Kautsar, S. A., Blin, K., Shaw, S., Navarro-Muñoz, J. C., Terlouw, B. R., van der Hooft, J. J. J., van Santen, J. A., Tracanna, V., Suarez Duran, H. G., Pascal Andreu, V., Selem-Mojica, N., Alanjary, M., Robinson, S. L., Lund, G., Epstein, S. C., Sisto, A. C., Charkoudian, L. K., Collemare, J., Linington, R. G., Weber, T., and Medema, M. H. (2019). MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Research*, **48**(D1), D454–D458.

Kaweewan, I., Komaki, H., Hemmi, H., and Kodani, S. (2017). Isolation and structure determination of new antibacterial peptide curacomycin based on genome mining. *Asian Journal of Organic Chemistry*, **6**(12), 1838–1844.

Kersten, R. D., Yang, Y.-L., Xu, Y., Cimermancic, P., Nam, S.-J., Fenical, W., Fischbach, M. A., Moore, B. S., and Dorrestein, P. C. (2011). A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nat. Chem. Biol.*, **7**(11), 794–802.

Leao, T., Castelão, G., Korobeynikov, A., Monroe, E. A., Podell, S., Glukhov, E., Allen, E. E., Gerwick, W. H., and Gerwick, L. (2017). Comparative genomics uncovers the prolific and distinctive metabolic potential of the cyanobacterial genus. *Proc. Natl. Acad. Sci. U. S. A.*, **114**(12), 3198–3203.

Medema, M. H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J. B., Blin, K., de Bruijn, I., Chooi, Y. H., Claesen, J., Coates, R. C., Cruz-Morales, P., Duddela, S., Düsterhus, S., Edwards, D. J., Fewer, D. P., Garg, N., Geiger, C., Gomez-Escribano, J. P., Greule, A., Hadjithomas, M., Haines, A. S., Helfrich, E. J. N., Hillwig, M. L., Ishida, K., Jones, A. C., Jones, C. S., Jungmann, K., Kegler, C., Kim, H. U., Kötter, P., Krug, D., Masschelein, J., Melnik, A. V., Mantovani, S. M., Monroe, E. A., Moore, M., Moss, N., Nützmann, H.-W., Pan, G., Pati, A., Petras, D., Reen, F. J., Rosconi, F., Rui, Z., Tian, Z., Tobias, N. J., Tsunematsu, Y., Wiemann, P., Wyckoff, E., Yan, X., Yim, G., Yu, F., Xie, Y., Aigle, B., Apel, A. K., Balibar, C. J., Balskus, E. P., Barona-Gómez, F., Bechthold, A., Bode, H. B., Borriss, R., Brady, S. F., Brakhage, A. A., Caffrey, P., Cheng, Y.-Q., Clardy, J., Cox, R. J., De Mot, R., Donadio, S., Donia, M. S., van der Donk, W. A., Dorrestein, P. C., Doyle, S., Driessen, A. J. M., Ehling-Schulz, M., Entian, K.-D., Fischbach, M. A., Gerwick, L., Gerwick, W. H., Gross, H., Gust, B., Hertweck, C., Höfte, M., Jensen, S. E., Ju, J., Katz, L., Kaysser, L., Klassen, J. L., Keller, N. P., Kormanec, J., Kuipers, O. P., Kuzuyama, T., Kyrpides, N. C., Kwon, H.-J., Lautru, S., Lavigne, R., Lee, C. Y., Linquan, B., Liu, X., Liu, W., Luzhetskyy, A., Mahmud, T., Mast, Y., Méndez, C., Metsä-Ketelä, M., Micklefield, J., Mitchell, D. A., Moore, B. S., Moreira, L. M., Müller, R., Neilan, B. A., Nett, M., Nielsen, J., O'Gara, F., Oikawa, H., Osbourn, A., Osburne, M. S., Ostash, B., Payne, S. M., Pernodet, J.-L., Petricek, M., Piel, J., Ploux, O., Raaijmakers, J. M., Salas, J. A., Schmitt, E. K., Scott, B., Seipke, R. F., Shen, B., Sherman, D. H., Sivonen, K., Smanski, M. J., Sosio, M., Stegmann, E., Süssmuth, R. D., Tahlan, K., Thomas, C. M., Tang, Y., Truman, A. W., Viaud, M., Walton, J. D., Walsh, C. T., Weber, T., van Wezel, G. P., Wilkinson, B., Willey, J. M., Wohlleben, W., Wright, G. D., Ziemert, N., Zhang, C., Zotchev, S. B., Breitling, R., Takano, E., and Glöckner, F. O. (2015). Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.*, **11**(9), 625–631.

Mohimani, H., Kersten, R. D., Liu, W.-T., Wang, M., Purvine, S. O., Wu, S., Brewer, H. M., Pasa-Tolic, L., Bandeira, N., Moore, B. S., Pevzner, P. A., and Dorrestein, P. C. (2014). Automated genome mining of ribosomal peptide natural products. *ACS Chem. Biol.*, **9**(7), 1545–1551.

Mohimani, H., Gurevich, A., Mikheenko, A., Garg, N., Nothias, L.-F., Ninomiya, A., Takada, K., Dorrestein, P. C., and Pevzner, P. A. (2017). Dereplication of peptidic natural products through database search of mass spectra. *Nat. Chem. Biol.*, **13**(1), 30–37.

Mohimani, H., Gurevich, A., Shlemov, A., Mikheenko, A., Korobeynikov, A., Cao, L., Shcherbin, E., Nothias, L.-F., Dorrestein, P. C., and Pevzner, P. A. (2018). Dereplication of microbial metabolites through database search of mass spectra. *Nat. Commun.*, **9**(1), 4035.

Navarro-Muñoz, J. C., Selem-Mojica, N., Mullowney, M. W., Kautsar, S. A., Tryon, J. H., Parkinson, E. I., De Los Santos, E. L. C., Yeong, M., Cruz-Morales, P., Abubucker, S., Roeters, A., Lokhorst, W., Fernandez-Guerra, A., Cappelini, L. T. D., Goering, A. W., Thomson, R. J., Metcalf, W. W., Kelleher, N. L., Barona-Gomez, F., and Medema, M. H. (2020). A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.*, **16**(1), 60–68.

Newman, D. J. and Cragg, G. M. (2020). Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J. Nat. Prod.*

Schorn, M., Verhoeven, S., van der Hooft, J., Medema, M., and Dorrstein, P. (2020). iOMEGA paired data platform. http://pairedomicsdata.bioinformatics.nl/. Accessed: 2020-3-21.

Soldatou, S., Eldjarn, G. H., Huerta-Uribe, A., Rogers, S., and Duncan, K. R. (2019). Linking biosynthetic and chemical space to accelerate microbial secondary metabolite discovery. *FEMS Microbiol. Lett.*, **366**(13).

van der Hooft, J. J. J., Mohimani, H., Bauermeister, A., Dorrestein, P. C., Duncan, K. R., and Medema, M. H. (2020). Linking genomics and metabolomics to chart specialized metabolic diversity. *Chem. Soc. Rev.*, pages –.

Wang, M., Carver, J. J., Phelan, V. V., Sanchez, L. M., Garg, N., Peng, Y., Nguyen, D. D., Watrous, J., Kapono, C. A., Luzzatto-Knaan, T., Porto, C., Bouslimani, A., Melnik, A. V., Meehan, M. J., Liu, W.-T., Crüsemann, M., Boudreau, P. D., Esquenazi, E., Sandoval-Calderón, M., Kersten, R. D., Pace, L. A., Quinn, R. A., Duncan, K. R., Hsu, C.-C., Floros, D. J., Gavilan, R. G., Kleigrewe, K., Northen, T., Dutton, R. J., Parrot, D., Carlson, E. E., Aigle, B., Michelsen, C. F., Jelsbak, L., Sohlenkamp, C., Pevzner, P., Edlund, A., McLean, J., Piel, J., Murphy, B. T., Gerwick, L., Liaw, C.-C., Yang, Y.-L., Humpf, H.-U., Maansson, M., Keyzers, R. A., Sims, A. C., Johnson, A. R., Sidebottom, A. M., Sedio, B. E., Klitgaard, A., Larson, C. B., P, C. A. B., Torres-Mendoza, D., Gonzalez, D. J., Silva, D. B., Marques, L. M., Demarque, D. P., Pociute, E., O'Neill, E. C., Briand, E., Helfrich, E. J. N., Granatosky, E. A., Glukhov, E., Ryffel, F., Houson, H., Mohimani, H., Kharbush, J. J., Zeng, Y., Vorholt, J. A., Kurita, K. L., Charusanti, P., McPhail, K. L., Nielsen, K. F., Vuong, L., Elfeki, M., Traxler, M. F., Engene, N., Koyama, N., Vining, O. B., Baric, R., Silva, R. R., Mascuch, S. J., Tomasi, S., Jenkins, S., Macherla, V., Hoffman, T., Agarwal, V., Williams, P. G., Dai, J., Neupane, R., Gurr, J., Rodríguez, A. M. C., Lamsa, A., Zhang, C., Dorrestein, K., Duggan, B. M., Almaliti, J., Allard, P.-M., Phapale, P., Nothias, L.-F., Alexandrov, T., Litaudon, M., Wolfender, J.-L., Kyle, J. E., Metz, T. O., Peryea, T., Nguyen, D.-T., VanLeer, D., Shinn, P., Jadhav, A., Müller, R., Waters, K. M., Shi, W., Liu, X., Zhang, L., Knight, R., Jensen, P. R., Palsson, B. O., Pogliano, K., Linington, R. G., Gutiérrez, M., Lopes, N. P., Gerwick, W. H., Moore, B. S., Dorrestein, P. C., and Bandeira, N. (2016). Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat. Biotechnol.*, **34**(8), 828–837.

Watrous, J., Roach, P., Alexandrov, T., Heath, B. S., Yang, J. Y., Kersten, R. D., van der Voort, M., Pogliano, K., Gross, H., Raaijmakers, J. M., Moore, B. S., Laskin, J., Bandeira, N., and Dorrestein, P. C. (2012). Mass spectral molecular networking of living microbial colonies. *Proc. Natl. Acad. Sci. U. S. A.*, **109**(26), E1743–52.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Model.*, **28**(1), 31–36.

Willighagen, E. L., Mayfield, J. W., Alvarsson, J., Berg, A., Carlsson, L., Jeliazkova, N., Kuhn, S., Pluskal, T., Rojas-Chertó, M., Spjuth, O., Torrance, G., Evelo, C. T., Guha, R., and Steinbeck, C. (2017). The chemistry development kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J. Cheminform.*, **9**(1), 33.

Xu, B., Aitken, E. J., Baker, B. P., Turner, C. A., Harvey, J. E., Stott, M. B., Power, J. F., Harris, P. W. R., Keyzers, R. A., and Brimble, M. A. (2018). Genome mining, isolation, chemical synthesis and biological evaluation of a novel lanthipeptide, tikitericin, from the extremophilic microorganism strain T81. *Chem. Sci.*, **9**(37), 7311–7317.