Coevolution-based prediction of protein-protein interactions in polyketide biosynthetic assembly lines

Bioinformatics

Wang, Yan; Correa Marrero, Miguel; Medema, Marnix H.; Dijk, Aalt D.J.

https://doi.org/10.1093/bioinformatics/btaa595

Bioinformatics, 36(19), 2020, 4846–4853
doi: 10.1093/bioinformatics/btaa595
Advance Access Publication Date: 27 June 2020
Original Paper

OXFORD

Sequence analysis

Coevolution-based prediction of protein–protein interactions in polyketide biosynthetic assembly lines

Yan Wang[1,†], **Miguel Correa Marrero**[1,‡], **Marnix H. Medema**[1,*] and **Aalt D. J. van Dijk**[1,2,*]

[1]Bioinformatics Groupand and [2]Department of Plant Sciences Biometris, Wageningen University & Research, 6708 PB Wageningen, The Netherlands

*To whom correspondence should be addressed.
†Present address: Department of Neurology, University Medical Center, Utrecht Brain Center, Utrecht University, Utrecht, The Netherlands
‡Present address: European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Cambridge, UK

Associate Editor: Alfonso Valencia
Received on November 17, 2019; revised on May 20, 2020; editorial decision on June 16, 2020; accepted on June 19, 2020

Abstract

Motivation: Polyketide synthases (PKSs) are enzymes that generate diverse molecules of great pharmaceutical importance, including a range of clinically used antimicrobials and antitumor agents. Many polyketides are synthesized by *cis*-AT modular PKSs, which are organized in assembly lines, in which multiple enzymes line up in a specific order. This order is defined by specific protein–protein interactions (PPIs). The unique modular structure and catalyzing mechanism of these assembly lines makes their products predictable and also spurred combinatorial biosynthesis studies to produce novel polyketides using synthetic biology. However, predicting the interactions of PKSs, and thereby inferring the order of their assembly line, is still challenging, especially for cases in which this order is not reflected by the ordering of the PKS-encoding genes in the genome.

Results: Here, we introduce PKSpop, which uses a coevolution-based PPI algorithm to infer protein order in PKS assembly lines. Our method accurately predicts protein orders (93% accuracy). Additionally, we identify new residue pairs that are key in determining interaction specificity, and show that coevolution of N- and C-terminal docking domains of PKSs is significantly more predictive for PPIs than coevolution between ketosynthase and acyl carrier protein domains.

Availability and implementation: The code is available on http://www.bif.wur.nl/ (under 'Software').

Contact: marnix.medema@wur.nl or aaltjan.vandijk@wur.nl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Polyketides are a structurally diverse group of specialized metabolites produced by polyketide synthases (PKS) in taxonomically diverse organisms. They have a large variety of natural functions, and many of them are used in the clinic as antibiotics, chemotherapeutics, immunosuppressants or cholesterol-lowering agents (Demain, 2014; Newman and Cragg, 2016). Among various polyketide producing systems, modular type I PKSs have attracted interest due to their unique catalytic mechanism: large multi-domain enzymes that consist of repeating modules that function in an assembly-line fashion to extend a growing polyketide chain until it is offloaded and often cyclized to form the core scaffold of the polyketide natural product (Dutta *et al.*, 2014; Fischbach and Walsh, 2006; Robbins *et al.*, 2016; Weissman, 2016). Each module in such an assembly line again consists of multiple domains. The three core domains are an acyltransferase (AT) domain that selects an extender unit, an acyl carrier protein (ACP) that functions as a 'hook' to which this unit is transferred by the AT domain and a ketosynthase (KS) domain that catalyzes the condensation reaction to extend the polyketide chain with this unit. Besides these three, PKSs may also contain ketoreductase (KR), dehydratase (DH) or enoyl reductase (ER) domains, which determine the degree to which an extender unit is reduced. According to the way AT domains function within the assembly lines, modular type I PKSs are divided into *cis*-AT PKSs, in which AT domains are present in all modules, and trans-AT PKSs, in which a standalone AT domain iteratively delivers extender units to each of the modules. In this study, we focus on *cis*-AT PKSs, which make up the majority (∼75%) of modular PKSs (Weissman, 2016).

PKS proteins may consist of one or multiple modules, and a PKS assembly line frequently consists of multiple individual PKS proteins, sometimes even 5–10 in total. The final structure of the polyketide is determined by the order in which the PKS proteins interact to form the assembly line (Dodge *et al.*, 2018). Therefore, understanding protein–protein interactions (PPIs) between PKSs is of great importance. After all, accurate prediction of these interactions from sequence information is essential for predicting the chemical structure of the product, and thus makes it possible to 'dereplicate' the biosynthetic gene clusters (BGCs) that encode the production of polyketides to identify which of these are responsible for the production of either known or unknown (and therefore novel) molecules (Dejong *et al.*, 2016). Moreover, understanding and predicting these protein interactions is crucial for synthetic biology approaches to PKS engineering to be successful (Klaus *et al.*, 2016), as it facilitates new ways of reordering PKS assembly lines to generate novel chemistry.

Several studies have attempted to predict the order of proteins in PKS assembly lines from sequence. The simplest way of doing this is by assuming collinearity of the proteins with the order of the PKS-encoding genes in a BGC, which has frequently been observed to be true for BGCs encoding the production of known molecules (Donadio *et al.*, 1991; Donadio and Katz, 1992; Yu *et al.*, 1999). However, there are many cases in which the order of the PKS-encoding genes in a BGC is not compatible with collinearity (e.g. when they are located in multiple operons) (Jørgensen *et al.*, 2009; Sun *et al.*, 2003; Takaishi *et al.*, 2013; Wenzel *et al.*, 2008; Zhang *et al.*, 2007), as shown in Figure 1. For example, in the MIBiG repository (Medema *et al.*, 2015), 16 out of 133 (12%) of the multimodular *cis*-AT PKS assembly lines are non-collinear. There are currently 3302 *cis*-AT PKS gene clusters in the antiSMASH database and the number increases with more and more genomes being sequenced in the future. This necessitates the use of more advanced approaches to predict PPIs. Several studies have shown that small helical domains at the N- and C-termini of PKS proteins, called docking domains, play a key role in mediating the specificity of interactions, through specific pairwise dimerization with docking domains of other proteins in the assembly line (Broadhurst *et al.*, 2003; Buchholz *et al.*, 2009; Tsuji *et al.*, 2001; Weissman, 2006b). Early computational analyses of the diversity of these docking domains showed that they fall into at least three compatibility classes (referred to as classes I-III, later renamed classes 1a, 1b and 2), and that head-to-tail pairing of docking domains within the same compatibility class is necessary but not sufficient to predict the order of PKS proteins in an assembly line (Thattai *et al.*, 2007; Yadav *et al.*, 2009). Additionally, several studies identified potential specificity-conferring (and coevolving) residue pairs, which made it possible to predict the final order of PKS assembly lines to a larger extent (Burger and van Nimwegen, 2008; Thattai *et al.*, 2007; Yadav *et al.*, 2009). Still, for the large majority of assembly lines, predictions of the latest (rule-based) methods have been inconclusive, as multiple orderings are deemed equally likely (Yadav *et al.*, 2009).

Excitingly, several developments have taken place in the past years that provide new opportunities to revisit and improve predictions of PKS assembly line ordering from sequence. First, new crystal structures of cyanobacterial ("class 2") docking domains (Whicher *et al.*, 2013) have revealed a markedly different tertiary structure for these domains, which shows that previous analyses based on joint alignments of multiple docking domain classes (Yadav *et al.*, 2009) may have been suboptimal. Second, considerable amounts of new data have been collected into standardized repositories (Medema *et al.*, 2015), making it possible to substantially expand the training sets. Third, we have recently developed Ouroboros, an algorithm that exploits the tendency of residues interacting across protein–protein interfaces to coevolve in order to maintain the interaction. Coevolution refers to the fact that interactions between amino acids which are important for protein structure and function impose constraints on the sets of mutations acceptable at interacting sites. In a multiple sequence alignment (MSA), this is reflected in statistical dependencies between different positions. Analysis of coevolution has recently gained momentum as an approach to predict protein contacts (de Juan *et al.*, 2013; Marks *et al.*, 2011), and Ouroboros exploits this idea towards prediction of PPIs. The algorithm models coevolution between MSAs of the proteins of interest to accurately distinguish between interacting and non-interacting proteins (Marrero *et al.*, 2018). Efforts to predict protein–protein information from coevolution have also been made by (Cong *et al.*, 2019), whose approach additionally uses protein structure information as one of the steps, while Ouroboros takes sequence information as the only input.

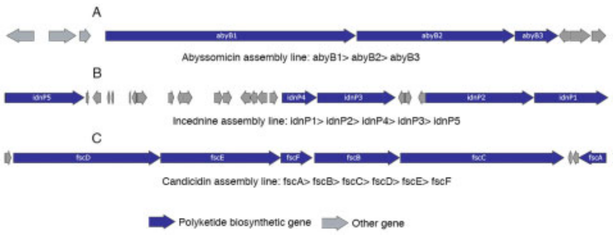
A
abyB1
abyB2
abyB3
Abyssomicin assembly line: abyB1> abyB2> abyB3
B
idnP5
idnP4
idnP3
idnP2
idnP1
Incednine assembly line: idnP1> idnP2> idnP4> idnP3> idnP5
C
fscD
fscE
fscF
fscB
fscC
fscA
Candicidin assembly line: fscA> fscB> fscC> fscD> fscE> fscF
Polyketide biosynthetic gene
Other gene

Fig. 1. Collinear and non-collinear PKS assembly lines. In each panel, the gene arrows represent the PKS biosynthetic gene cluster (BGC); the protein order is provided underneath. (**A**) The protein order of the abyssomicin assembly line is collinear with the order of genes in the genome. (**B**) The genes in the incednine BGC span three operons. The order of the proteins in the assembly line is non-collinear with the order of the genes within the operons. (**C**) In the candicidin BGC the genes span three operons. Although there is collinearity within each operon, the order of operons is not collinear with the order of proteins in the assembly line

Here, we provide a new prediction pipeline, PKSpop, that exploits these new insights and developments, through automated recognition of docking domain classes, coevolutionary prediction of interaction probabilities between class 1a docking domains using expectation maximization (EM) with the Ouroboros method (Marrero *et al.*, 2018), and combining these results in a new algorithm that is usually able to predict a single most probable assembly line order using only sequence data. We show that PKSpop is able to accurately identify assembly line orders in non-collinear PKS systems in 93% of the cases. We compare PKSpop with two previously presented approaches and demonstrate clear improvement in prediction performance. Additionally, we identify new residue pairs that are key in determining interaction specificity, and show that docking domain coevolution is significantly more predictive for PPIs than coevolution between ACP domains and KS–AT linker regions.

2 Materials and methods

2.1 Dataset for learning and evaluation

Sequences of interacting docking domain pairs were obtained from The Minimum Information about a Biosynthetic Gene cluster (MIBiG) repository (Medema *et al.*, 2015). A total of 372 interacting docking domain pairs were extracted from 133 *cis*-AT PKS assembly lines, whose protein orders have been established by previous studies. In addition, 15 interacting pairs were obtained from 13 assembly lines in the antiSMASH database, which stores the BGCs detected and annotated by antiSMASH (Blin *et al.*, 2017b; Weber *et al.*, 2015); these were all assembly lines consisting of two or three proteins, for which there was only one possible order based on the positioning of loading module (a module comprising only an AT domain and an ACP domain) (Weissman and Müller, 2008) and/or thioesterase domain at the start and end. An extra set of docking domain pairs was extracted from adjacent genes in gene clusters in the antiSMASH database. The interaction status of these pairs is unknown.

Compared to the ACP structure (Alekseyev *et al.*, 2007), the ACP domains annotated in the MIBiG and antiSMASH databases did not cover the whole ACP sequence. Thus, the sequences detected by antiSMASH as ACP domains were extended with 20 additional N-terminal residues. 'KS–AT linker' regions refer to the sequences between KS and AT domains, which have been reported to contact their upstream ACP domains during intermodule and interprotein polyketide chain translocation (Kapur *et al.*, 2010, 2012). In addition to the interprotein KS–AT linker/ACP pairs, the intra-protein pairs, whose KS–AT linker and ACP domain were on the adjacent module in a protein, were extracted.

Datasets of 384 docking domain pairs and 1287 KS–AT linker/ACP pairs were obtained after being filtered by removing redundant sequences at 100% identity, as identified by CD-HIT (Li and Godzik, 2006). CD-HIT was also used to perform an additional experiment in which redundant sequences were filtered at a 90% identity cut-off.

2.2 Clustering and multiple sequence alignment

Profile HMM analysis (Eddy, 1998) was employed to classify and align the docking domains. It should be noted that the naming of compatibility classes I, II and III that refer red to docking domain sequences (Thattai *et al.*, 2007) has been updated as described by (Weissman, 2016): sequence compatibility classes I, II and III correspond to structural classes 1a, 1b and 2, respectively. In our study, N-terminal sequences (sequences before the first KS domain in proteins) and C-terminal sequences (sequences after the last ACP domain in proteins) of different compatibility classes published by (Thattai *et al.*, 2007) were aligned separately by MUSCLE (Edgar, 2004) with a gap-opening penalty of 11.0. Using the alignment, hmmbuild from the HMMER package (Eddy, 2011) was used to build HMM profiles for each class, and hmmscan was then used to assign class membership to sequences by finding for each sequence its best match among the three HMM profiles. Subsequently, the sequences belonging to different classes were aligned against their profile by hmmalign. The conserved regions, 16 residues at Class 1a and 1b C-terminal sequences, 24 residues at Class 2 C-terminal sequences and 26 residues at N-terminal sequences, which are involved in the PPI (Broadhurst *et al.*, 2003, 2009; Weissman, 2006a; Whicher *et al.*, 2013), were obtained from the MSAs as docking domains (Supplementary Fig. S1). Docking domains used in study are described in Supplementary Datasets S1 and S2. For various alignments, the number of effective sequences (N_{eff}) was calculated with conkit v0.11.2 (Simkovic *et al.*, 2017). The KS–AT linkers and ACP domains were aligned by MUSCLE with a gap-opening penalty of 10.0, respectively.

2.3 Ouroboros analysis

In order to maintain a PPI across evolution, interface residues coevolve. By modeling coevolution across interacting proteins, we can identify these interacting residues, but it is non-trivial to avoid adding pairs of non-interacting proteins into the analysis, which add noise. We previously developed Ouroboros (Marrero *et al.*, 2018), an algorithm that models coevolution to distinguish between pairs of interacting and non-interacting proteins in two families and thereby better identify interacting residues. Based on the detection of a coevolutionary signal, i.e. statistical dependencies between columns in MSAs, the method iterates between inferring PPI probabilities and predicting correlated mutations between alignments taking into account these PPI probabilities. By doing this, we can filter out non-interacting proteins from the alignments, thereby boosting the coevolutionary signal and improving contact prediction performance. The method requires no training labels regarding interactions nor contacts.

Datasets with different percentages of interacting pairs and different numbers of effective sequences were analyzed with Ouroboros. Since Ouroboros employs EM, each analysis was repeated three times using different random seeds to address the problem that EM might find a local optimum. In addition, different values of the int_frac parameter were used. Out of these analyses, the result with the largest log-likelihood was selected. Each dataset could consist of interacting pairs, non-interacting pairs and pairs without interaction information; the assessment of performance of PPI prediction was always based on the known interacting and non-interacting pairs.

2.4 Selecting specificity-determining residues

Ouroboros predicts contacts by assigning each residue pair a contact score. In the structure of interacting class 1a docking domain pair (PDB # 1PZR), there are 31 physical contact residue pairs under the threshold of 5 Å (Thattai *et al.*, 2007). Therefore, the residue pairs of the top 31 contact scores were considered as the Ouroboros-predicted contacts and the correct prediction was defined as the intersection between the Ouroboros prediction and the physical contacts. Note that the structural information is only used for validation and interpretation of the results, not as input for the prediction: PKSpop uses the sequence of the whole domain (and not just the specificity-determining residues) to predict the protein order.

To find the set of residues that define the specificity of PKS PPIs, we checked the predictive performance of Ouroboros in the absence of each residue pair separately and removed the columns from the MSA corresponding to the residue pair that had the least impact on PPI prediction. With the remaining residues, we again removed the pair that had the least impact. This procedure was repeated until all residue pairs were removed. The performance was assessed by the AUC value of the ROC curve, calculated by scikit-learn v0.20.2 (Varoquaux *et al.*, 2015). PyMOL v2.0 (pymol.org) was used to visualize the position of selected residues in the structure.

2.5 Logistic regression model

The docking domain pairs and KS-AT linker/ACP pairs were analyzed with Ouroboros separately. Then, a logistic regression model was built to predict the interaction of proteins that harbour class 1a docking domains. One of the predictors was the interaction probability of docking domain pairs predicted by Ouroboros; the other predictor was the interaction probability of the corresponding KS–AT linker/ACP from the same protein pair. The dataset contained 80% interactions and 20% non-interactions. Extra docking domain pairs from adjacent genes in gene clusters and intra-protein KS-AT linker/ACP pairs were added to the datasets to increase sequence diversity and were not used to examine prediction performance. Five-fold cross-validation was used to test the performance of the logistic regression model.

2.6 PKSpop: protein order prediction pipeline

A pipeline, PKSpop, was designed and implemented to predict the order of interactions in PKS assembly lines. PKSpop takes the GenBank file of the query PKS gene cluster as detected by antiSMASH and a list of protein identifiers of the PKSs in that gene cluster as input. The code is available on http://www.bif.wur.nl/ (under 'Software').

PKSpop starts by extracting the docking domain sequences from the PKS proteins encoded in the cluster. First, the start protein of the assembly line is determined by locating a loading module (AT–ACP); the end protein is determined by locating a thioesterase domain. The N-terminal sequences (sequences before the first KS domain in proteins) and C-terminal sequences (sequences after the last ACP domain in proteins) are then assigned to one of the three different docking domain classes by hmmscan (using competitive scoring of class-specific docking domain pHMMs), and aligned using hmmalign (see Section 2.2). Proteins whose N-terminus or C-terminus was not assigned to any class are recognized as start or end protein. If there is more than one start or end protein in the cluster, the pipeline stops and outputs 'No prediction'.

The conserved 26-amino acid and 16-amino acid regions in the class 1a N- and C-termini are subsequently obtained from the alignment as N-terminal docking domains (Ndds) and C-terminal docking domains (Cdds). Next, each of the Ndds is paired with all Cdds. Two paired fasta files, one containing Ndds and one containing Cdds, are generated by integrating the query sequences with 222 known interacting sequences and 301 sequences from adjacent genes in gene clusters. The two fasta files with paired docking domain sequences are then input into Ouroboros and analyzed under user-defined int_frac parameters (default [0.80, 0.90]) and a user-indicated number of times the analysis is repeated. Note that the interaction information of the internally integrated 523 sequence pairs is not used by Ouroboros.

Ouroboros then outputs the pairwise interaction probabilities (0–1) of the query docking domain pairs from the result with the largest log-likelihood. These are filled into an interaction probability matrix, where the Cdds are in rows and the Ndds are in columns. If there are class 1b/2 termini, the interaction probabilities of the class

1b/2 pairs are assigned 1. Due to the compatibility class and assembly line constraints, interaction should not happen between (i) the start protein (defined by encoding a loading module or lacking a Ndd) and any Cdds, (ii) the end protein (defined by having a thioesterase domain or lacking a Cdd) and any Ndds, (iii) the Cdd and Ndd on the same protein and (iv) docking domains that belong to different classes. Interaction probabilities of these combinations are therefore assigned 0. Then, probabilities (Probs) are transformed to weights (-Probs). Interacting pairs are selected from the matrix by the Hungarian algorithm (Kuhn, 2005) [implemented by SciPy v1.2.1 (Millman *et al.*, 2011)], which finds a match of Cdds to Ndds in which the sum of weights is minimum. Pairs that are in conflict with assembly line constraints are removed from the selected ones. The protein order is predicted based on the remaining pairs.

3 Results & discussion

3.1 A new method to predict PKS order in assembly line

Predicting protein order in polyketide assembly lines is different from the typical pairwise PPI studies. This is because of the nature of the assembly line, in which a protein can only interact with one upstream and one downstream protein. Before presenting the prediction results, we first provide an overall overview of PKSpop, the pipeline developed to apply coevolutionary analysis to PKS assembly lines. To infer protein order in an assembly line, PKSpop comprises three main steps: (i) Identify class memberships for query docking domains and align the sequences. (ii) Pair each class 1a N-terminal docking domain (Ndd) with all class 1a C-terminal docking domains (Cdds) and use coevolution to predict the interaction probability for all possible pairs and enter these into a probability matrix. (iii) Infer protein order from the probability matrix by the Hungarian algorithm, a global optimization method, which finds a match between Ndds and Cdds that maximizes the overall interaction probability. The inference method takes the assembly line constraints and compatibility class into account; in particular, the fact that the first and last domain in the assembly line can be recognized from the sequence allows converting predicted PPIs to a predicted protein order for the entire assembly line.

We first set out to automatically classify docking domains into classes, and to assess for which of these classes sufficient data are available to predict their PPIs. The performance of coevolution-analysis-based PPI prediction depends highly on the alignment quality, and shifts in interaction site can lead to mispredictions (Marrero *et al.*, 2018; Uguzzoni *et al.*, 2017). While a previous method (Yadav *et al.*, 2009) used joint alignment of the docking domains from all three compatibility classes (Thattai *et al.*, 2007), new structural studies have shown that these docking domain classes have a markedly different tertiary structure and therefore should not be aligned (Broadhurst *et al.*, 2003; Buchholz *et al.*, 2009; Whicher *et al.*, 2013). Instead, it is better to analyze the classes separately. To annotate docking domains by class, we built HMM profiles of all three docking domain compatibility classes from sequences published by (Thattai *et al.*, 2007). Then, 384 interacting Cdd and Ndd pairs, which were selected from MIBiG and the antiSMASH database, were clustered by finding for each domain its best match among the three HMM profiles. As a result, 222 class 1a and 85 class 1b interacting docking domain pairs were obtained. There were 64 pairs that contained at least one class 2 docking domain; 27 of these involved cases where a single class 2 docking domain was paired with a domain which was not assigned to any class. There were 13 pairs in which both docking domains were not assigned to any class. One possible reason is that only 14 sequences were used to build the class 2 HMM profile which might be too little to cover the sequence diversity of class 2 docking domains. The other possible reason might be that class 2 was not correctly defined: in previous clustering (Thattai *et al.*, 2007) (where it was called 'class III'), most pairs that were ostensibly mismatched based on comprising members of different classes have one of their sequences assigned into class 2, and, in the sequence similarity networks, class 2 appears to be much less coherent then classes 1a and 1b (Supplementary Methods, Fig. S2). Therefore, it is still an open question whether there is a sufficiently homogeneous 'class 2' that truly represents a group of structurally similar docking domains across organisms.

Using the alignments of the class 1a docking domains, we applied Ouroboros, our recently developed method to PPIs based on intermolecular coevolution. Ouroboros takes only the alignments as input and generates pairwise protein interaction probabilities of the class 1a docking domains. Testing various input datasets, we demonstrate that coevolutionary analysis can be successfully applied to predict class 1a docking domain interactions.

3.2 PKSpop predicts PKS protein orders with high accuracy

To assess the performance of the PKSpop pipeline specifically in the context of assembly lines that do not follow the collinearity rule, predictions were performed on 14 non-collinear *cis*-AT PKS assembly lines independently and then evaluated. As a result, PKSpop achieved a 93% accuracy at predicting the protein order in 14 non-collinear *cis*-AT assembly lines and 98% accuracy at detecting 55 of 59 true PPIs from the 336 possibly interacting protein pairs in these assembly lines (Supplementary Dataset S3). Considering that there are *n*! possible orders for an assembly line that comprises *n* proteins, the probability of obtaining a 93% accuracy on predicting the protein order by random chance is 1.5e-27. Figure 2 shows an example of predicting the protein order of the candicidin assembly line. FscA is the first protein, as it has an initiating AMP-binding domain, and FscF is the final protein as it does not have a Cdd (Fig. 2A). The Cdd of FscD (FscD-Cdd) and Ndd of FscE (FscE-Ndd) were clustered into class 1b, while other docking domains all belong to class 1b. The pairwise interaction probabilities of the docking domains were predicted by Ouroboros and filled into the matrix (Fig. 2B). The pair without an interaction probability was assigned 0 or 1 probability based on the compatibility class the docking domains belonged to and the assembly line constraints. The probability matrix was then transformed to a weight matrix. The Hungarian algorithm was applied to find a match of Cdds to Ndds that minimizes the sum of weights (maximize the overall interaction probability). Subsequently, the protein order was predicted from the matched docking domain pairs.

To further investigate the performance of PKSpop, it was then applied to 50 collinear assembly lines in the MIBiG repository. In this analysis, PPIs were predicted with 95% accuracy; the resulting accuracy for predicting assembly lines was 80%. Among the mispredicted assembly lines, BGC0000086 (MIBiG accession) and BGC0000149 contain more than one protein that do not have an Ndd. In BGC0001051, TgaC was recognized as both the start protein and the end protein and was hence removed from the query proteins. The order of the remaining proteins was predicted. It indicates that the PPIs of the above three assembly lines differ from the generally observed PPI in *cis*-AT PKS system, which is via the interaction of docking domains. Leaving them out of the test set and taking the collinear and non-collinear assembly lines together, PKSpop predicts PPIs with 97% accuracy, and assembly lines with 87% accuracy (Supplementary Dataset S3).

PKSpop performs better with shorter assembly lines: a 91% accuracy is achieved on predicting the assembly lines of ≤6 proteins, while 40% accuracy is obtained on those of >6 proteins. This is related to the fact that the complexity of prediction increases factorially with the increase of the number of proteins. The probability of correctly predicting an assembly line that comprises n proteins is 1/*n*! by chance. For example, the probability of correctly predicting an assembly line of seven proteins by chance is 1/5040, which is seven times smaller than that of a six-protein assembly line (1/720). In addition, all the possibly interacting protein pairs in an assembly line have to be used as input for Ouroboros to infer interaction probability. This means that the longer the assembly line, the more non-interacting pairs are introduced, which will reduce the performance of Ouroboros (discussed in Section 3).

We compared the performance of our algorithm to the previously published method of (Yadav *et al.*, 2009), which uses two residue pairs as a 'docking code' to infer protein orders. With this method, only 42% of the assembly lines in our test sets were correctly predicted (Supplementary Dataset S3). Moreover, it should be noted

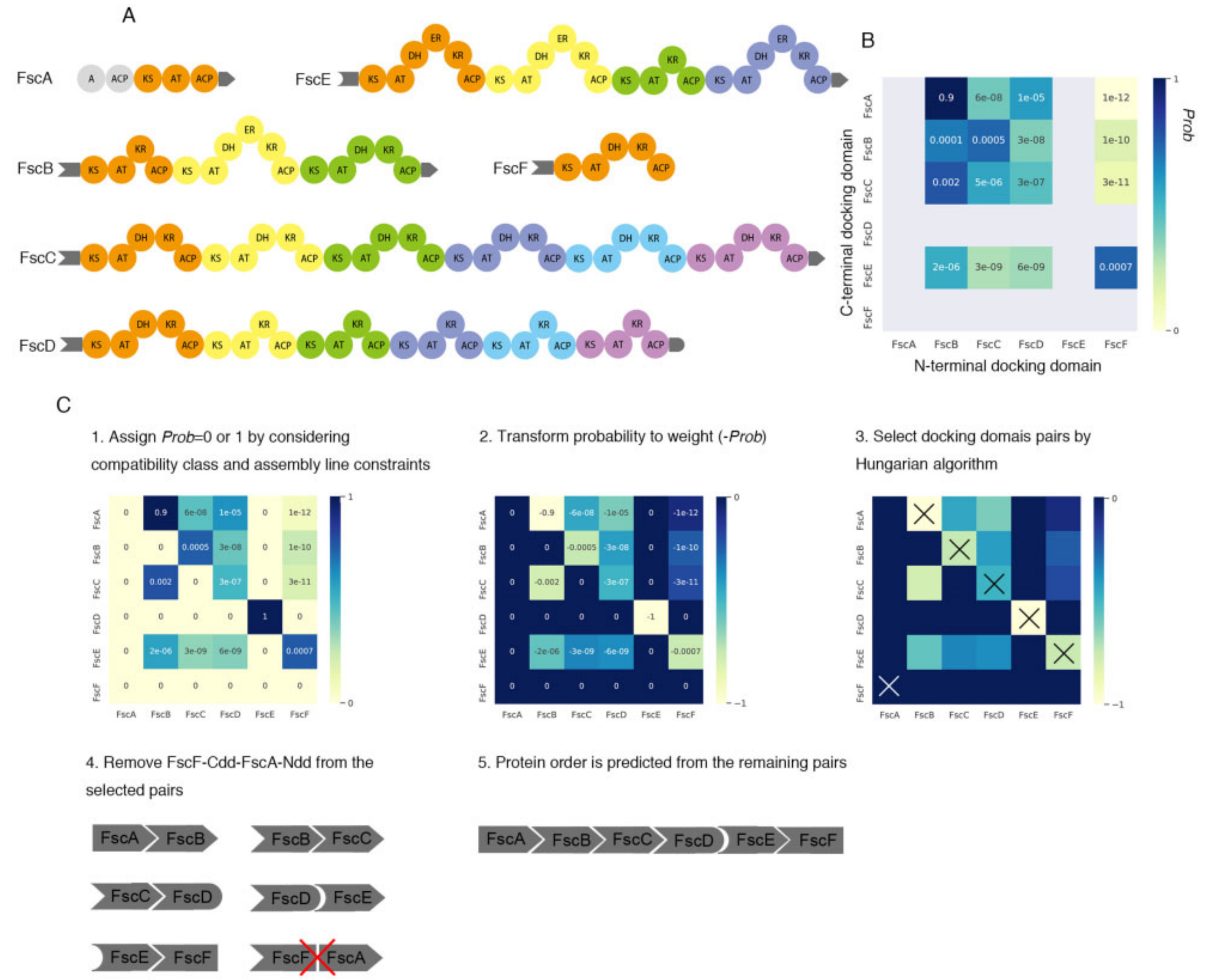
A
FscA
A ACP KS AT ACP
FscB
FscC
FscD
FscE
FscF
KS AT DH KR ER ACP
B
C-terminal docking domain
N-terminal docking domain
FscA FscB FscC FscD FscE FscF
0.9 6e-08 1e-05 1e-12
0.0001 0.0005 3e-08 1e-10
0.002 5e-06 3e-07 3e-11
2e-06 3e-09 6e-09 0.0007
Prob
C
1. Assign Prob=0 or 1 by considering compatibility class and assembly line constraints
2. Transform probability to weight (-Prob)
3. Select docking domais pairs by Hungarian algorithm
4. Remove FscF-Cdd-FscA-Ndd from the selected pairs
5. Protein order is predicted from the remaining pairs
FscA FscB
FscB FscC
FscC FscD
FscD FscE
FscE FscF
FscF FscA
FscA FscB FscC FscD FscE FscF

Fig. 2. Predicting the protein order of Candicidin assembly line. (**A**) The PKS proteins and their domains in the Candicidin assembly line. (**B**) Interaction of PKSs predicted by Ouroboros. Cells coloured by the value of interaction probability (Prob) represent a possible interaction of a docking domain pair. Cells coloured in grey represent docking domain pairs without an interaction probability: FscE-Ndd and FscD-Cdd do not belong to class 1a; FscA (FscF) are the start (end) protein without Ndd (Cdd). (**C**) The interaction probability of FscD-Cdd and FscE-Ndd is assigned 1, as they comprise the pair that does not belong to class 1a. The interactions of class 1a docking domains to FscD-Cdd and FscE-Ndd are all assigned 0 probability. Due to the PPI constraints in the assembly line, the FscA-Ndd column, FscF-Cdd row and the diagonal line in the matrix are assigned 0. Then, probabilities are transformed to weights (-Probs). The Hungarian algorithm applied to the transformed matrix selects six docking domain pairs; of these, FscF-Cdd-FscA-Ndd is removed, as FscA contains the loading module and should therefore not be preceded by another protein in the assembly line. The remaining pairs provide the predicted protein order

that our method outputs a single order for each assembly line. In contrast, the previous method predicted multiple possible orders for each assembly line (with a maximum of 36 possibilities for the correctly predicted assembly lines). This further underlines the advantage of our approach. A recently published machine learning-based tool, DDAP, predicts multiple orders ranked by likelihood for an assembly line (Li *et al.*, 2019). In their study, evaluation was performed using cross-validation on 89 assembly lines; for 71% of the cases, the true order ranked among the top three. Note that we cannot directly compare DDAP and PKSpop on the non-collinear assembly lines for which we perform predictions, since these cases were all included as training data for DDAP. Compared to DDAP, PKSpop performed better in predicting a single correct order. Moreover, the fact that PKSpop makes use of coevolution to predict interactions, means that it can be applied to study the molecular basis of interaction specificity (discussed in the next section). This is not directly possible with the machine learning-based approach of DDAP.

Besides the adopted Hungarian algorithm, which finds the globally best match of Cdds to Ndds, a greedy method which selects docking domain pairs with the highest probability iteratively from the interaction matrix to predict the protein order was also developed (Supplementary Methods, Figs S3 and S4). It had an 86% accuracy on predicting the non-collinear assembly lines and 85% accuracy on the collinear assembly lines, which is still better than the previous method but underperformed with respect to the global optimization strategy used by the Hungarian algorithm.

3.3 Identification of specificity-conferring residue pairs in class 1a docking domains

In addition to inferring PPIs, Ouroboros is able to predict the contact residues that mediate PPIs based on coevolutionary signal. Based on structural information, there are 31 contacting residue pairs on class 1a docking domains. We selected the top 31 residue pairs scored by Ouroboros as involved in intermolecular coevolution, and of these, 21 were involved in intermolecular contacts (*P*-value 7.47e-15, Fisher exact test), showing that the method can indeed find such contacts.

In order to obtain a better understanding of how these selected residue pairs impact PPI specificity, MSAs were generated in which all the predicted residues were removed. The prediction of PPIs based on the remaining residues failed (AUC = 0.55), which

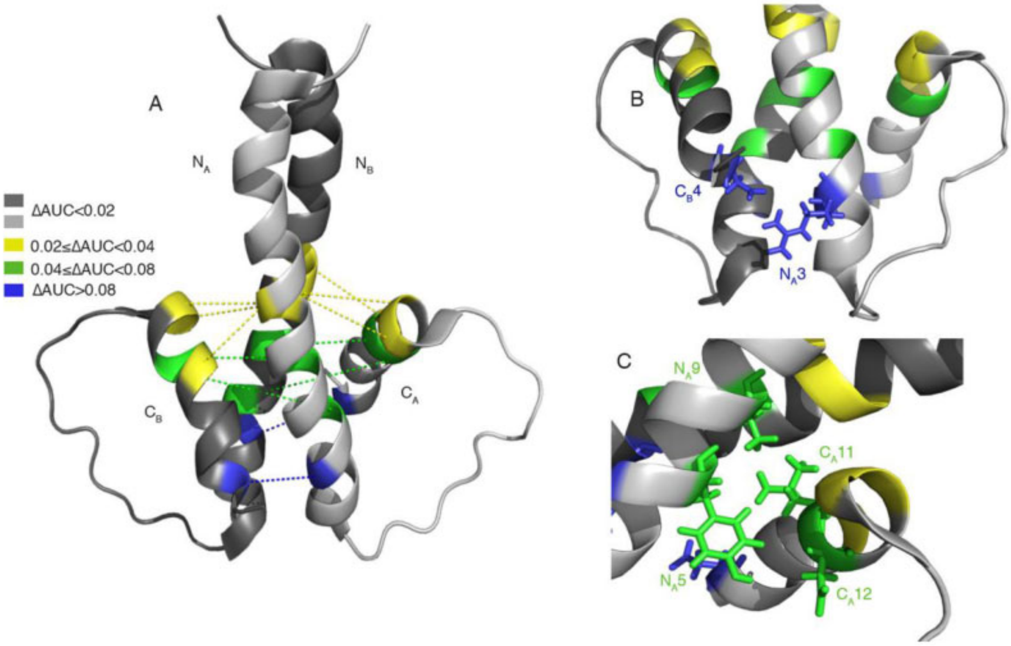

Fig. 3. Class 1a docking domain interaction structure with predicted crucial residues. (**A**) The docking domain interaction structure (PDB # 1PZR). PKS proteins are homodimers and therefore there are two docking domain pairs, CA-NA and CB-NB. Residue pairs (linked by dashed lines) are coloured according to the decrease of AUC after they were removed from the sequences. (**B**) The physical position of residue pair C4-N3, which was predicted to contribute to protein interaction specificity by Ouroboros. (**C**) The physical position of C12-N5 and C11-N9, which were also predicted to contribute to protein interaction specificity. The C12-N5 pair was found for the first time in our study as a determinant of PPI specificity

indicates that the determinant residue pairs lie in the selected pairs. The knowledge of residues that determine PPI specificity can facilitate engineering synthetic PKS assembly lines. To figure out which residue pairs are most predictive of PPIs, a recursive feature elimination was done on the 21 selected pairs. By removing the least important contact residue pairs one by one from the MSA, we found that performance decreased drastically after pairs C11-N9, C12-N5 and C4-N3 were removed (Supplementary Fig. S5). Removing all three pairs still led to a model with more predictive power (AUC = 0.67) than a model in which all contacting pairs were removed. However, when using the three pairs alone, the predictive performance (AUC = 0.83) is similar to that of using the whole sequence (AUC = 0.81) (Supplementary Table S1). This indicates that they are likely to be specificity-determining residues (Fig. 3).

Next, we analyzed how the three specificity-determining residue pairs influence the order prediction (Supplementary Table S1). Using the three pairs to infer protein order led to an accuracy of 85%, which is slightly less than the accuracy (87%) resulted from using the whole sequence. It indicates that the three pairs together explain most of PPI specificity. Predicting protein order with each of the three pairs alone resulted in relatively high accuracy (79% for C11-N9, 77% for C4-N3, 74% for C12-N5), which suggests that they were successfully identified as the specificity-determining residues. While C4-N3 and C11-N9 were also recognized as the protein-contacting parts of the ‘code word’ of PPI specificity by (Thattai *et al.*, 2007), the C12-N5 pair is found for the first time in our study as a determinant of PPI specificity.

3.4 Docking domains instead of KS–AT linker/ACP pairs contribute to predicting PPI

Besides docking domains, the regions between KS domains and AT domains, which were referred to as KS–AT linkers, have also been shown to physically contact upstream ACP domains during intermodule polyketide chain translocation and play a role in mediating PPIs between PKSs (Dutta *et al.*, 2014; Kapur *et al.*, 2010; Klaus *et al.*, 2016). To investigate whether KS–AT linker/ACP pairs can also contribute to predicting PPI specificity in PKSs, a logistic regression model was used to predict the interaction of PKS proteins that harbour a class 1a docking domain. One of the predictors was the interaction probability of docking domain pairs predicted by Ouroboros. The other predictor was the interaction probability of the corresponding KS–AT linker/ACP pairs. Different training and testing sets were created by 5-fold cross-validation, and the ROC curve was plotted on each testing set (Supplementary Fig. S6). The resulting AUC of 0.84 ± 0.04 demonstrates that the model is predictive of PKS–protein interactions. In the logistic model, the coefficient of the KS–AT linker/ACP interaction probability was 0.39 ± 0.50, much lower than that of docking domain, 2.87 ± 0.24, which indicates that the Ouroboros' result on KS–AT linker/ACP pairs adds only limited value to the PPI prediction. To further evaluate this result, interaction probabilities of the two pairs obtained from Ouroboros were used separately to predict the PPI (Fig. 4). The ROC curves of docking domains have similar shape and AUCs with that of logistic regression model, while the model based on KS–AT linker/ACP is only slightly better than a random guess.

The poor predictive ability of Ouroboros result on KS–AT linker/ACP pairs might indicate that coevolution is not suitable to predict KS–AT linker/ACP interaction. According to previous studies (Kapur *et al.*, 2010, 2012), R23 on DEBS2 ACP helix I was found to be the determinant of inter-protein KS–AT linker/ACP interaction specificity by mutagenesis (Kapur *et al.*, 2012), but the contact score of this residue to any other KS–AT linkers residues is relatively low in the Ouroboros result. The low contact score might be explained by the high conservation of R23 (R in 70% of the ACP sequences). The poor predictive ability might also indicate that KS–AT linker/ACP interaction is not involved in mediating the specificity of PPIs between PKSs. Circumstantial evidence for this is given by the DEBS assembly line, for which functional protein interaction only requires compatible docking domain pairs (Dodge *et al.*, 2018; Wu *et al.*, 2002).

3.5 Noise level and sequence variation influence the prediction

Predicting the protein order in assembly lines is based on predictions of PPIs by Ouroboros, and this prediction can be influenced by the noise level (number of non-interacting protein pairs) and sequence variation of the input proteins (Marrero *et al.*, 2018). The impact

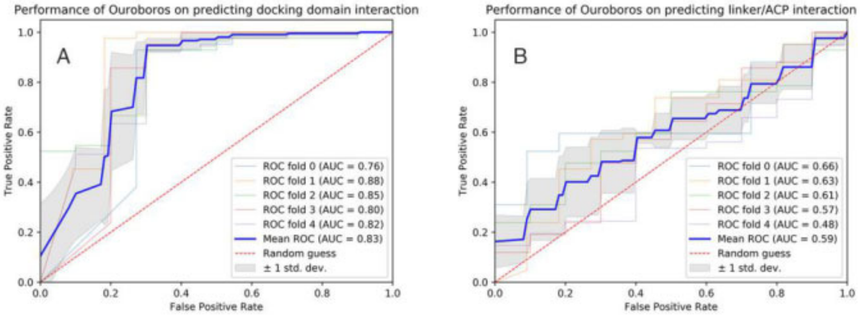

Fig. 4. PKS protein interaction prediction performance. Receiver operating characteristic (ROC) curves of logistic models based on inferred interaction probabilities, using only (**A**) class 1a docking domains and (**B**) KS-AT linker/ACP to predict the PPIs

Table 1. Predictive performance with different percentages of interacting pairs and different number of effective sequences

Dataset composition: % of interacting pairs	N_{eff} (<90% identity)	MCC
60%	212 ± 1	0.22 ± 0.02
60% + extra pairs	468 ± 4	0.45 ± 0.06
80%	211 ± 3	0.30 ± 0.05
80% + extra pairs	465 ± 2	0.60 ± 0.07

Note: Mean values of 3 (with 60% interacting pairs) or 5 (with 80% interacting pairs) datasets.

that variation in these factors has on prediction performance will be different for different protein families, given that these vary in characteristics such as the number of available sequences. To investigate the dependency of our method on the number of PKS sequences, we tested Ouroboros with different docking domain datasets.

To examine the performance of Ouroboros in predicting PPI, datasets of non-interacting domain pairs were generated by pairing docking domains from non-interacting proteins within each PKS assembly line. To compare the predictive performance with different input noise levels (number of non-interacting protein pairs), five datasets with 80% interacting class 1a docking domains and three datasets with 60% interacting domains were generated. All datasets comprised the same 222 interacting domain pairs and a number of different non-interacting domain pairs randomly picked from the non-interacting datasets. Sequence variation was measured by the number of effective sequences (N_{eff}, number of sequences in a dataset whose pairwise identities are below 90%). To increase N_{eff}, extra docking domain pairs were added to the datasets of 80% and 60% interacting pairs. The 301 extra pairs were generated by pairing the docking domains from adjacent genes in polyketide BGCs compatible with collinearity. According to the collinearity rule, these extra data are likely to include a high percentage of interacting pairs and can be used to increase the sequence variation without introducing too much noise. The interaction information of these extra pairs is unknown, and the performance assessment was based on the known interacting and non-interacting pairs. Importantly, the knowledge on interaction status was not used for training the method, only for evaluating its performance. Table 1 shows Ouroboros' predictive performance, measured by the Matthew correlation coefficient, on the different datasets. The results clearly demonstrate that with more interacting domain pairs and greater N_{eff} in the dataset, Ouroboros tends to perform better. Therefore, when using Ouroboros to generate the pairwise interaction matrix in the order prediction pipeline, the query sequences are combined with all the existing interacting protein pairs and the extra pairs.

Since Ouroboros is able to infer interaction of proteins that contain class 1a docking domains with good accuracy (Supplementary Fig. S7C), it was then applied to class 1b docking domains. Lacking equal amounts of sequence data, an MSA with a N_{eff} (<90% identity) of only 168 was created, which comprised a set of 80% interacting pairs and a set of extra pairs without interaction information. The result shows that Ouroboros failed to perform well on the class 1b docking domains with this amount of effective sequences (Supplementary Fig. S7). To analyze whether the poor performance is caused by the lack of sequence variation, datasets of class 1a docking domain with similar N_{eff} (<90% identity) were created and input into Ouroboros. The performance on class 1a domains is indeed better, but the standard deviation is relatively high compared to that obtained with the higher N_{eff}. Altogether, this suggests that the performance is unstable on such a small dataset. In order to assess specifically the impact of the number of sequences with high mutual sequence identity in the dataset on the predictive performance, five datasets were generated by filtering out sequences above 90% identity in the class 1a docking domain datasets of $465 \pm 2\, N_{\text{eff}}$. As shown in Supplementary Fig. S7C and D, there was only a small decrease of PPI prediction performance after removing the sequences of high identity. This is in line with the fact that for a coevolution-based approach, what matters is sequence diversity: very similar sequences will not add much information.

With more PKS BGCs being sequenced in the future, there will be more class 1b docking domains available, which can increase the sequence variation in MSA and may help improve the prediction for this class.

The fact that currently Ouroboros cannot accurately predict protein interactions for class 1b or class 2 docking domains means that PKSpop is currently unable to infer assembly lines whose docking domains all belong to class 1b or 2. However, as demonstrated above, the PKSpop pipeline can work on assembly lines that contain limited numbers of class 1b or 2 docking domain pairs (in addition to class 1a docking domain pairs) by connecting the docking domains of class 1b and/or 2, respectively. Based on the composition of docking domain classes in BGCs currently deposited in MIBiG, we estimate that PKSpop is able to predict assembly line orders in ~90% of the cases.

4 Conclusions and future perspectives

We have presented a method to predict protein order in PKS biosynthetic assembly lines. Our method, PKSpop, enables us to deal with the fact that often, gene order in a gene cluster is not predictive of the protein interaction order in the assembly line. The basis of our approach is our recently developed coevolution-based method for protein interaction prediction, which assigns a likelihood of interaction to pairs of proteins based on whether coevolution is predicted between these proteins. We adopted the Hungarian algorithm to convert the coevolution-based interaction probability between multiple proteins to an ordering of proteins in an assembly line. Application of this approach to a set of known PKS gene clusters demonstrated its performance. To predict the ordering, our method does not require any knowledge on protein interactions or residue contacts. Structure information was only used as background knowledge to define the protein domains on which the analysis could focus, and was used to interpret the predicted interacting residue pairs in a structural context.

PKSpop now enables to predict PPIs and hence putative core scaffold chemical structures for clusters which defy the collinearity rule; hitherto, interactions in these clusters could not be well predicted. In the near future, we plan to incorporate PKSpop as a prediction feature into the antiSMASH pipeline (Blin *et al.*, 2017a).

In addition, our method identified specific residue pairs, one of which was highlighted for the first time as specificity determining; this pair might be further studied to provide insight into the determinants of PKS specificity. Finally, in contrast to collinearity-based predictions, our approach enables predicting interactions between proteins from different assembly lines. This will be of great use for efforts to engineer synthetic PKS assembly lines (Hertweck, 2015; Poust *et al.*, 2014; Weissman, 2016) consisting of entirely new combinations of proteins.

Financial Support: none declared.

Conflict of Interest: none declared.

References

Alekseyev,V.Y. *et al.* (2007) Solution structure and proposed domain-domain recognition interface of an acyl carrier protein domain from a modular polyketide synthase. *Protein Sci.*, **16**, 2093–2107.

Blin,K. *et al.* (2017a) antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.*, **45**, W36–W41.

Blin,K. *et al.* (2017b) The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res.*, **45**, D555.

Broadhurst,R.W. *et al.* (2003) The structure of docking domains in modular polyketide synthases. *Chem. Biol.*, **10**, 723–731.

Buchholz,T.J. *et al.* (2009) Structural basis for binding specificity between subclasses of modular polyketide synthase docking domains. *ACS Chem. Biol.*, **4**, 41–52.

Burger,L. and van Nimwegen,E. (2008) Accurate prediction of protein–protein interactions from sequence alignments using a Bayesian method. *Mol. Syst. Biol.*, **4**, 165.

Cong,Q. *et al.* (2019) Protein interaction networks revealed by proteome coevolution. *Science*, **365**, 185–189.

Dejong,C.A. *et al.* (2016) Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. *Nat. Chem. Biol.*, **12**, 1007–1014.

Demain,A.L. (2014) Importance of microbial natural products and the need to revitalize their discovery. *J. Ind. Microbiol. Biotechnol.*, **41**, 185–201.

Dodge,G.J. *et al.* (2018) Protein–protein interactions in '*cis*-AT' polyketide synthases. *Nat. Prod. Rep.*, **35**, 1082–1096.

Donadio,S. et al. (1991) Modular organization of genes required for complex polyketide biosynthesis. *Science*, **252**, 675–679.

Donadio,S. and Katz,L. (1992) Organization of the enzymatic domains in the multifunctional polyketide synthase involved in erythromycin formation in *Saccharopolyspora erythraea*. *Gene*, **111**, 51–60.

Dutta,S. *et al.* (2014) Structure of a modular polyketide synthase. *Nature*, **510**, 512–517.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**, e1002195.

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

Fischbach,M.A. and Walsh,C.T. (2006) Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. *Chem. Rev.*, **106**, 3468–3496.

Hertweck,C. (2015) Decoding and reprogramming complex polyketide assembly lines: prospects for synthetic biology. *Trends Biochem. Sci.*, **40**, 189–199.

Jørgensen,H. *et al.* (2009) Biosynthesis of macrolactam BE-14106 involves two distinct PKS systems and amino acid processing enzymes for generation of the aminoacyl starter unit. *Chem. Biol.*, **16**, 1109–1121.

de Juan,D. *et al.* (2013) Emerging methods in protein co-evolution. *Nat. Rev. Genet.*, **14**, 249–261.

Kapur,S. *et al.* (2010) Molecular recognition between ketosynthase and acyl carrier protein domains of the 6-deoxyerythronolide B synthase. *Proc. Natl. Acad. Sci. USA*, **107**, 22066–22071.

Kapur,S. *et al.* (2012) Reprogramming a module of the 6-deoxyerythronolide B synthase for iterative chain elongation. *Proc. Natl. Acad. Sci. USA*, **109**, 4110–4115.

Klaus,M. *et al.* (2016) Protein-protein interactions, not substrate recognition, dominate the turnover of chimeric assembly line polyketide synthases. *J. Biol. Chem.*, **291**, 16404–16415.

Kuhn,H.W. (2005) The Hungarian method for the assignment problem. *Naval Res. Logist.*, **52**, 7–21.

Li,T. *et al.* (2020) DDAP: docking domain affinity and biosynthetic pathway prediction tool for type I polyketide synthases. *Bioinformatics, 36, 942-944.*

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Marks,D.S. *et al.* (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS One*, **6**, e28766.

Marrero,M.C. *et al.* (2019) Improved inference of intermolecular contacts through protein–protein interaction prediction using coevolutionary analysis. *Bioinformatics, 35, 2036–2042.*

Medema,M.H. *et al.* (2015) Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.*, **11**, 625–631.

Millman,K.J. *et al.* (2011) Python for scientists and engineers. *Comput. Sci. Eng.*, **13**, 9–12.

Newman,D.J. and Cragg,G.M. (2016) Natural products as sources of new drugs from 1981 to 2014. *J. Nat. Prod.*, **79**, 629–661.

Poust,S. *et al.* (2014) Narrowing the gap between the promise and reality of polyketide synthases as a synthetic biology platform. *Curr. Opin. Biotechnol.*, **30**, 32–39.

Robbins,T. *et al.* (2016) Structure and mechanism of assembly line polyketide synthases. *Curr. Opin. Struct. Biol.*, **41**, 10–18.

Simkovic,F. *et al.* (2017) ConKit: a python interface to contact predictions. *Bioinformatics*, **33**, 2209–2211.

Sun,Y. *et al.* (2003) A complete gene cluster from Streptomyces nanchangensis NS3226 encoding biosynthesis of the polyether ionophore nanchangmycin. *Chem. Biol.*, **10**, 431–441.

Takaishi,M. *et al.* (2013) Identification of the incednine biosynthetic gene cluster: characterization of novel β-glutamate-β-decarboxylase IdnL3. *J. Antibiot.*, **66**, 691–699.

Thattai,M. *et al.* (2007) The origins of specificity in polyketide synthase protein interactions. *PLoS Comput. Biol.*, **3**, 1827–1835.

Tsuji,S.Y. *et al.* (2001) Selective protein-protein interactions direct channeling of intermediates between polyketide synthase modules. *Biochemistry*, **40**, 2326–2331.

Uguzzoni,G. *et al.* (2017) Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis. *Proc. Natl. Acad. Sci. USA*, **114**, E2662–E2671.

Varoquaux,G. *et al.* (2015) Scikit-learn. *GetMobile Mobile Comput. Commun.*, **19**, 29–33.

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

Weissman,K.J. (2006a) Single amino acid substitutions alter the efficiency of docking in modular polyketide biosynthesis. *Chembiochem*, **7**, 1334–1342.

Weissman,K.J. (2006b) The structural basis for docking in modular polyketide biosynthesis. *Chembiochem*, **7**, 485–494.

Weissman,K.J. (2016) Genetic engineering of modular PKSs: from combinatorial biosynthesis to synthetic biology. *Nat. Prod. Rep.*, **33**, 203–230.

Weissman,K.J. and Müller,R. (2008) Protein–protein interactions in multienzyme megasynthetases. *ChemBioChem*, **9**, 826–848.

Wenzel,S.C. *et al.* (2008) A type I/type III polyketide synthase hybrid biosynthetic pathway for the structurally unique ansa compound kendomycin. *Chembiochem*, **9**, 2711–2721.

Whicher,J.R. *et al.* (2013) Cyanobacterial polyketide synthase docking domains: a tool for engineering natural product biosynthesis. *Chem. Biol.*, **20**, 1340–1351.

Wu,N. *et al.* (2002) Quantitative analysis of the relative contributions of donor acyl carrier proteins, acceptor ketosynthases, and linker regions to intermodular transfer of intermediates in hybrid polyketide synthases. *Biochemistry*, **41**, 5056–5066.

Yadav,G. *et al.* (2009) Towards prediction of metabolic products of polyketide synthases: an in silico analysis. *PLoS Comput. Biol.*, **5**, e1000351.

Yu,T.-W. *et al.* (1999) Direct evidence that the rifamycin polyketide synthase assembles polyketide chains processively. *Proc. Natl. Acad. Sci. USA*, **96**, 9051–9056.

Zhang,H. *et al.* (2007) Elucidation of the kijanimicin gene cluster: insights into the biosynthesis of spirotetronate antibiotics and nitrosugars. *J. Am. Chem. Soc.*, **129**, 14670–14683.