Machine learning to predict protein-protein interactions in polyketide biosynthetic assembly lines

MSc thesis
Bioinformatics Group

Yan Wang

Supervisors:
Marnix Medema
Aalt-Jan van Dijk

November 2018

Abstract

Polyketide synthases (PKSs) are multienzymes arranged in assembly lines that generate diverse polyketides of great pharmaceutical importance. The unique modular structure and catalyzing mechanism of these assembly lines makes their products predictable and also triggered the combinatorial biosynthesis study on producing novel polyketide drugs. These studies rely on the prediction of PKS protein-protein interaction (PPI) and the knowledge of residues that contribute to the interaction specificity. Most of the previous studies use only the docking domains to predict PKS PPI, without involving other domains. Here we adapt Ouroboros, an algorithm based on correlated mutation analysis, to predict PPI by analyzing both docking domains and KS-ACP domains interactions and identify the specificity determinant residues on the interaction surfaces. We also present an example of predicting PKS protein order in an assembly line.

Introduction

The synthesis of polyketides has always been of great interest given the pharmaceutical importance of polyketides (PK), which are used in clinic as antibiotics, antihyperlipidaemics, immunosuppressants and anticancer agents [1]. Many polyketides are synthesized by enzyme complexes of type I polyketide synthases organized in assembly lines. A module in an assembly line introduces one monomer unit to the polyketide chain, modifies its structure, and translocates the chain to the next module [2][3]. Each catalytic activity is discretely conducted by a domain in the module. A polyketide synthase (PKS) protein consists of one or more modules, and therefore, the order of proteins in an assembly line determines the structure of the final molecule product. Predicting PKS protein-protein interactions (PPI) in assembly lines can discover the order of proteins. Meanwhile, the monomer introduced by each module can be predicted by the sequence of AT domain [4]. Therefore, the chemical structure of polyketides can be predicted base on sequence information. Prediction of residues that define the interaction specificity also helps for creating novel PKS assembly lines using combinatorial biosynthesis strategies, which enables the design of novel polyketide drugs [5][35].

There are two short polypeptides at the C and N terminal of each PKS protein holding the assembly line together, referred to as docking domains. The C-terminal domain of the upstream protein binds to the N-terminal domain of the downstream protein. A previous study clustered C-terminal and N-terminal docking domains respectively into three phylogenetic clusters according to sequence similarity [6]. They found that a cluster of C-terminal domains generally only interacts with a corresponding cluster of N-terminal domains. There were still some interacting domain pairs from different compatibility classes and many non-interacting pairs from the same class. Another study extracted two crucial interacting residue pairs by structural alignment of various docking domains. The two pairs of residues were ranked as favorable, neutral, and unfavorable by residue charge and hydrophobicity. The predicted order of proteins in one assembly line should be the one that have the highest number of favorable residue pairs in the interfaces of adjacent proteins. The correct prediction rate of this simple approach was about 80%, which suggested that the two contact residue pairs can be used to predict the PPI. However, as there were many orders having the same high score, especially in multi-protein assembly lines, the correct prediction referred to eliminating a large proportion of low-score orders and narrowing down the correct order to a small number of combinatorial possibilities [7].

The KS-ACP interaction plays a vital role in both PK chain translocation and elongation. A study attempted to insert a RAPS (rapamycin synthases) module into DEBS (erythromycin synthases) 1, between module 1 and 2. The product of this hybrid protein showed that the RAPS module was skipped during the chain extension [8]. A separate study further showed that the intermediate was transferred from ACP(DEBS) to ACP(RAPS) instead of from ACP(DEBS) to KS(RAPS) [9]. The skipping might due to the weak interaction between hybrid ACP and KS domain.

Another study created a series of di-modular assembly line with a loading didomain (LDD), a DEBS module 1 (DEBS M1) and a variable module, and also created tri-modular assembly lines with a LDD, a DEBS M1, a variable module and a DEBS M3. The variable modules were derived from heterologous PKS. There were low or no yields of the hybrid lines and the yields of tri-module were lower than the di-module. This result was likely due to the poor substrate recognition of KS and weak interaction of ACPn-KSn+1. It has been shown in many studies that keeping ACPn-KSn+1 interface intact can lead to higher yields. This suggested that ACPn-KSn+1 interaction is crucial to chain translocation, and therefore responsible for the productivity of a hybrid line [36]. The region on ACP involved in intermodular recognition by KS has been studied through mutagenesis. Kapur S *et al*. created chimeric ACP domains from ACP2 and ACP4, and the chain transfer rates of the chimeras to KS3, which interact with ACP2 in natural assembly line, were compared [10]. They found that the transfer rates were higher when the helix I region was from ACP2, which means that the chain transfer could be controlled by ACP helix I. Then, they identified the residues that control chain transfer by mutating residues of ACP4 helix I to ACP2 residues one by one. As a result, residue 23 was found to contribute to the chain translocation, and another 3 putative contacting residues were found on computational structure [11]. Although the ACP interface involved in ACPn-KSn+1 interaction has been studied, the contacting residues on KS domains has not been found [12][32]. Also, previous studies applied PPI prediction algorithm only on the docking domains to predict PKS PPI, without involving KS-ACP [33][34].

Here we predict the PKS PPI by combining the interaction status of docking domains and KS-ACP with the hypothesis that the functional interaction of two PKS proteins should result from interactions of both docking domains and KS-ACP domains. The interaction of the two domain pairs are analyzed by Ouroboros, an algorithm predicting PPI by correlated mutation analysis [13].

Methods

Dataset

The data of interacting domain pairs comes from The Minimum Information about a Biosynthetic Gene cluster (MIBiG) [14]. 372 interacting docking domain pairs were extracted from 139 type 1 PKS assembly lines in the MIBiG. There are also 15 interacting pairs from 13 assembly lines in the antiSMASH database, which stores the biosynthetic gene clusters detected and annotated by antiSMASH [15][16]. The docking domains of unknown interaction status were also extracted, as they can increase the coevolution signal in the multiple sequence alignment.
Compared to the ACP physical structure [17], the ACP domains in the MIBiG and antiSMASH database did not cover the whole ACP sequences. Thus, the sequences detected by antiSMASH plus 20 residues ahead were extracted as ACP domains. Except for the interprotein KS/ACP pairs, the intraprotein pairs were also extracted to increase sequence variation. Sequences of extreme length were trimmed and the datasets were filtered by removing the redundant sequences of 100% identity identified by CD-HIT [18].

Clustering and multiple sequence alignment

Profile HMM analysis [19] was employed to align the docking domains. Sequences of different compatibility classes published by Thattai *et al* [6] were aligned separately by MUSCLE [20] with gap open penalty 11.0. *hmmbuild* from HMMER package was used to build HMM profiles for each class with the alignment. *hmmscan* was used to assign class membership to sequences by finding each sequence its best match among 3 HMM profiles. Then, the sequences belong to different classes were aligned against their profile by *hmmalign*. The conserved region of 26 residues on N-terminal docking domain and 16 residues at the end of C-terminal docking domains, where the protein-protein interaction happens [21][27][28], were obtained from the MSAs (Supplement figure 1). The short regions were then analyzed to predict PPI.
The collinearity of the order of PKS proteins and the order of encoding genes were found in many PKS biosynthesis

assembly lines [22]. Therefore, two extra datasets were generated by pairing the docking domains from the adjacent genes (301 class 1 docking domain pairs in a dataset, 189 class 2 docking domain pairs in the other dataset). The datasets are likely to have high percentage of interacting pairs and can be used to increase the sequence variation to the query sequences without introducing too much noise.

The type 1 PKS KS domains are different in modular PKS, hybrid PKS and *trans*-AT PKS [23][24][25]. As the hybrid KS were not included in the initial dataset, the modular KS were selected by matching KS domains to the modular-KS and *cis*-KS HMM profiles from antiSMASH (https://bitbucket.org/antismash/antismash/src/f32e28078b4a2c98a71453b6d9fed74707ccdb88/antismash/specific_modules/nrpspks/ksdomains.hmm?at=master&fileviewer=file-view-default). The ACPs paired with the modular KS were then selected.

Analysis of domain pairs by Ouroboros

Datasets of docking domains with different percentage of interacting pairs and different number of effective sequences were performed by Ouroboros. Since Ouroboros employs Expectation-Maximization (EM), each analysis was repeated three times using different random seeds for different values of the *int_frac* parameter to address the problem that EM will find a local optimum. The results presented are from the analysis that has the largest log likelihood. 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.

Logistic regression model

The docking domain pairs and KS-ACP pairs from the query protein pairs were analyzed by Ouroboros separately. Then, logistic regression model was performed to predict PKS PPI. One of the predictors is the interaction probability of docking domain pairs predicted by Ouroboros. The other predictor is the interaction probability of the corresponding KS-ACP pairs. The dataset contained 80% interactions and 20% noninteractions. The datasets were then split to 5 training sets and 5 testing sets to build and test the model by 5-fold cross validation. Here we used logistic regression model to predict the interaction of proteins that harbor class 1 docking domain.

Predicting residue contact

In the PDB structure of interacting class 1 docking domain pair (PDB: 1PZR), there are 31 physical contact residue pairs under the threshold of 5 Angstrom [6]. Ouroboros predicts contacts by assigning each residue pair a contact score and there is no threshold to define contact. Therefore, the residue pairs of top 31 contact scores were considered as the Ouroboros-predicted contacts and the correct prediction was defined as the intersection between Ouroboros prediction and physical contacts.

Selecting interaction specificity determinant residues

To find the set of residues that define the specificity of PKS PPI, we checked the predictive performance with the absence of each residues pair separately and removed one pair that has the least impact on PPI prediction from the MSA. With the remaining residues, again removed the one pair that has the least impact. This procedure was repeated until all residue pairs were removed.

Result & Discussion

HMM profiles of three docking domain compatibility classes were built from sequences published by Thattai *et al* [6]. Then, C-terminal and N-terminal docking domains were clustered respectively by finding for each domain its best match among the 3 HMM profiles. As a result, less than 5% of class 1 docking domains have a compatible domain that belong to other class, and that figure of class 2 is 8%. However, for almost half of the class 3 docking domains the compatible domains were not assigned to any class. One possible reason is that the number of sequences used to build class 3 HMM profiles was only 14, which might be too little to reflect on the characteristic of class 3. If this is the real case, the problem could be addressed after more class 3 docking domain are sequenced and clustered. The other possible reason might be that the "class 3" was not correctly defined. In Thattai's result [6], most of pairs mismatched in different classes have a sequence clustered into class 3. Also, the docking domains clustered in class 3 by the article do not have the similar physical structure, such as, the CurK/CurL (PDB: 4MYY) and CurG/CurH (PDB: 4MYZ) [26], which makes the class 3 difficult to be aligned.

According to their structural information [21][27], docking domain pairs from different compatibility classes are likely to have different interaction pattern. Thus, class 1 and class 2 docking domains were analyzed separately.

Ouroboros results on KS-ACP pairs does not contribute to predicting PPI

Logistic regression model was built to predict the interaction of PKS proteins that contain class 1 docking domain. One of the predictors is the interaction probability of docking domain pairs predicted by Ouroboros. The other predictor is the interaction probability of the corresponding KS-ACP pairs. Different training and testing sets were created by 5-fold cross-validation, and the ROC curve was plotted on each testing set (Figure 1). With average AUC equals to 0.83 and ±1 standard deviation area does not overlap with the random guess line, the figure suggests that the model is predictive to the PPI of PKSs.

In the logistic model, the coefficient of the KS-ACP interaction probability is around -0.29, much lower than that of docking domain, 2.98, which indicates that the Ouroboros result on KS-ACP does not contribute to the prediction. To evaluate this result, interaction probabilities of each domain pair obtained from Ouroboros were used separately to predict the PPI, showed in Figure 2. The ROC curves of docking domains have similar shape and AUCs with that of logistic regression model, while that of KS-ACP are not better than random guess.

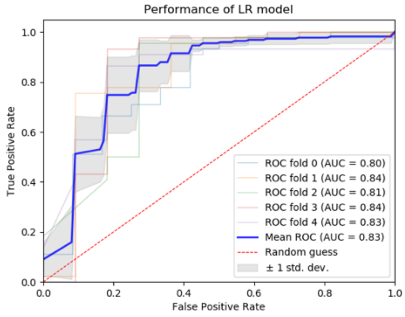


Figure 1. Performance of logistic regression model built on Ouroboros results of class 1 docking domain pairs and the corresponding KS-ACP pairs. The light curves show performances of model on 5 testing sets; the bold blue curve presents the averaged performance; the grey area is the estimated ROC ±1 standard deviation.

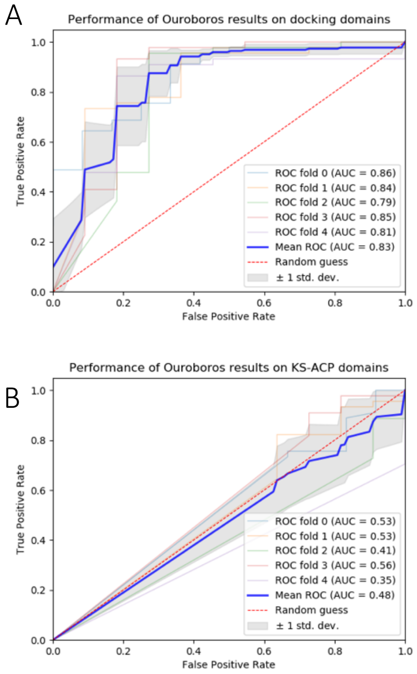


Figure 2. Performance of using only Ouroboros result on A) class 1 docking domains and B) KS-ACP to predict PPI. The domain pairs that plotted by each light curve are one of the two predictors behind the same line in the Figure 1.

The poor predictive ability of Ouroboros result on KS-ACP pairs might result from that 1) Ouroboros is not suitable to predict KS-ACP interaction, or 2) KS-ACP interaction is not involved in the PPI as assumed. According to previous studies [10][11], KS mainly interacts with the inter-protein ACP helix I region in substrate chain translocation and interacts with the intra-protein ACP loop I in chain elongation. To not confuse the Ouroboros, the ACP regions other than helix I were trimmed in the MSA. It is possible that there are other crucial residues located on the trimmed region. But the prediction was still not better than random guess when using ACP whole sequence in Ouroboros. R23 on DEBS2 ACP helix I was found to be the determinant of inter-protein KS-ACP interaction by mutagenesis [10], but the contact score of this residue to any other KS residues are relatively low in Ouroboros result. The low contact score could be explained by that the contact residue in KS domains did not mutate in the evolution. It is also possible that KS-ACP interaction does not determine the PPI, as there is case in DEBS that the protein interaction only needs compatible docking domain pairs [29][30].

Higher proportion of interacting protein pairs in dataset leads to better performance of Ouroboros

Since Ouroboros-predicted interaction probability of KS-ACP is not predictive, docking domains were analyzed as the only predictor of PPI. The noise level can greatly influence the Ouroboros performance [13]. 5 datasets with 80% interacting class 1 docking domains and 3 datasets with 60% interacting domains were generated. All the datasets comprised the same interacting domains set and a number of different non-interacting domain pairs. Figure 3A and 3B compares the performance of predicting PPI on datasets with different percentage of interacting docking domains. It clearly shows that with more interacting domain pairs in the dataset, Ouroboros tends to be more predictive.

Increasing the effective sequences improves the PPI prediction

The sequence variation in MSA, which is measured by the number of effective sequences (N_{eff}), also influence the Ouroboros performance. The influence is determined by the core of the algorithm, correlated mutation analysis [13][31]. To increase the N_{eff}, extra docking domain pairs were extracted from the antiSMASH database and added to the datasets of 80% and 60% interacting docking domain pairs. The interaction information of these extra pairs is unknown, so the performance assessment was still based on the interacting and non-interacting pairs (Figure 3C, 3D). Comparing to figure 3A and 3B, the performance was greatly improved. From the figures and Table 1, It seems that the performance on 60% interaction datasets improved more than that on 80% interaction datasets, with AUC increased 21% versus 13% and MCC increased 105% versus 70%. It could be explained by that there are more than 60% interacting pairs after adding extra pairs to the datasets.

Table 1. Predictive performance on datasets with different percent of interacting pairs and different effective sequences. The N_{eff} and Matthew correlation coefficient (MCC) are the mean values of 3 (with 60% interacting pairs) or 5 (with 80% interacting pairs) datasets. The last column refers to the MCC values when setting the threshold of interaction probability to 0.5.

Dataset composition: % of interacting pairs (+ extra pairs)	N_{eff}	MCC
60%	274	0.22
60% + extra pairs	473	0.45
80%	202	0.33
80% + extra pairs	404	0.56

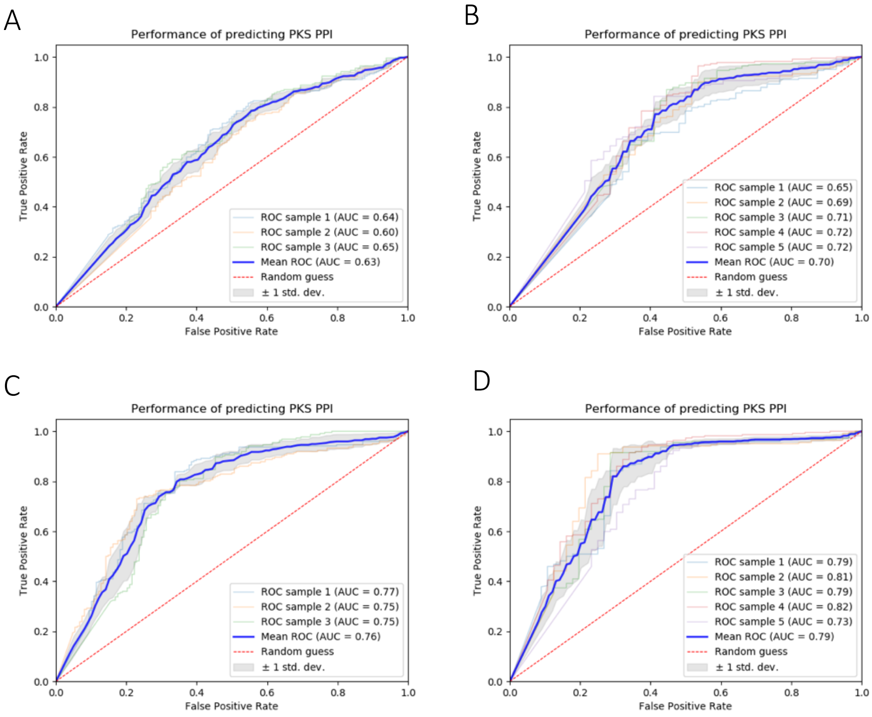


Figure 3. Predictive performance in dataset with different N_{eff} and different percentage of interacting docking domain pairs. Top two figures: ROC curves of Ouroboros-predicted interaction probability on datasets with A) 60% of interacting docking domains and B) 80% of interacting docking domains. Bottom two figures: ROC curves of Ouroboros results on C) 60% interacting pairs plus extra pairs datasets and D) 80% interacting pairs plus the same extra pairs datasets. In each figure, the datasets that each thin line plots consist of the same interacting docking domain pairs but different noninteracting pairs. Each light curve plots the Ouroboros result with the best LLH over different settings and repeats to one specific dataset. The bold blue curve presents the average performance and the grey area is the estimated ROC ±1 standard deviation.

Since Ouroboros is able to infer interaction of proteins that contain class 1 docking domains with reasonable accuracy, it was then applied to class 2 docking domains. Lacking equal amounts of sequence data, an MSA of only 168 N_{eff} 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 on the class 2 docking domains with this amount of effective sequences (Figure 4A). To analyze whether the poor performance is caused by the lack of sequence variation, datasets of class 1 docking domain with similar N_{eff} were created and input into Ouroboros. Figure 4 shows that the performance on class 1 domains is indeed better, but the standard deviation is relatively high compared to that in Figure 3. It suggests that Ouroboros is unstable on such small dataset. As there are only one dataset of class 2 domains, it is possible that the prediction would be better on different dataset. With more PKS biosynthetic gene clusters being sequenced in the future, there will be more class 2 docking domains available, which can increase the sequence variation in MSA and may help the prediction.

Increasing the proportion of interacting docking domain pairs and the N_{eff} can improve the prediction. Therefore, when predicting PKS PPI in practical, it would be better to combine the query sequences pairs with all the interacting docking domain pairs in MiBIG and the pairs from adjacent genes in antiSMASH.

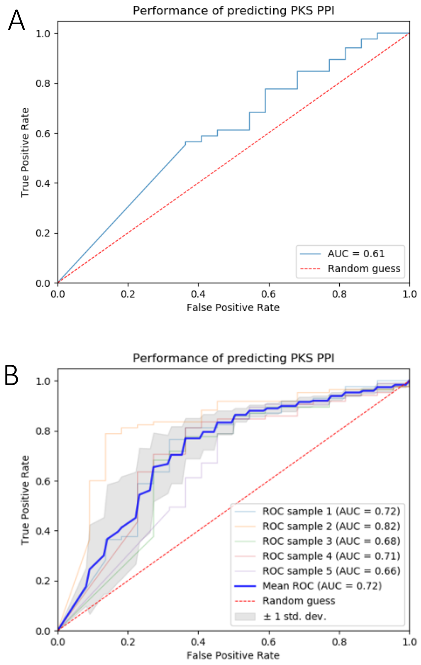


Figure 4. Predictive performance on one dataset of A) class 2 docking domains and five datasets on B) class 1 docking domains with similar N_{eff}.

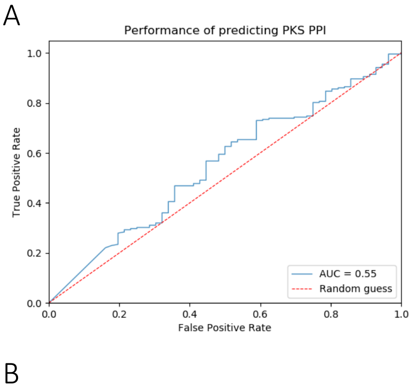
Residue pairs that are crucial to the PPI specificity

Ouroboros contact prediction is also influenced by the noise level and sequence variation. Datasets of different percent of interacting pairs and N_{eff} were fed to Ouroboros to compare contact prediction performance. The prediction results are showed in table 2. There are at least 13 correctly predicted contact pairs out of 31 pairs of highest contact scores (p-value 7.97e-7, Fisher exact test), which means the contact prediction for all datasets is not by chance.

Table 2. Summary of the number of correctly predicted contacts using different input datasets.

Dataset composition: % of interacting pairs (+ extra pairs)	Nr. correctly predicted contacts
100%	13
100% + extra pairs	13
80%	21
80% + extra pairs	15
60% + extra pairs	14

To study whether the residue pairs predicted by Ouroboros have impact on PPI specificity, MSAs without all predicted residues were generated and input into Ouroboros. The prediction result shows that the prediction failed without these residues (Figure 5A), which indicates that the determinant residue pairs lies in these pairs. The predictive performance decreased drastically when pair {4,3} was removed, with AUC dropped from 0.81 (Figure 3D) to 0.69 (Figure 5B), which suggest that {4,3} contributes to the PPI specificity. {4,3} was also identified by a previous study as the specificity determinant residue pair [21]. To find other residues that contribute to the interaction specificity, contact pairs that predicted by Ourorboros were removed one by one from the MSA (the last section in Method). The performance decreased after {11,9}, {12,5} and {4,3} were removed from the MSA (Supplement figure 2R, 2T, 2U). But removing {11.9} alone seems not influence the prediction and removing {12,5} alone has some influence (Supplement figure 3A, 3B), which suggest that the contact residues might have a combinatorial influence on the interaction. In other words, using all relevant residues is much more predictive than using only the specificity determinant residues. {5,1} and {14,14} were identified as "specificity code" by Yadav G *et al* [7]. But when they were removed from the MSA, the prediction performance did not decrease (Supplementary Figure 3C, 3D). This might result from that, in Yadav's work, they aligned different classes of docking domains together, ignoring the different positions of contacting residues in each class.



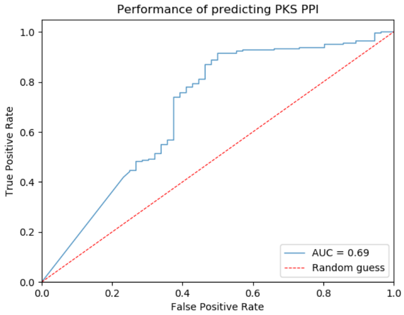


Figure 5. Predictive performance with A) the absence of all the residue pairs predicted to contact by Ouroboros and B) the absence of residue pair {4,3}.

Example of predicting the protein order in an assembly line

To show how Ouroboros can be used to predict the protein order in a PKS assembly line, a type 1 PKS biosynthetic gene cluster in antiSMASH database was analyzed (Genome *Streptomyces sp*. 769, NZ_CP003987, Gene Cluster 17, Figure 6). In this cluster, there is a TE domain in the GZL_RS16750 and no N-terminal docking domain in GZL_RS16785, which indicates that they are the last and the first PKS in the assembly line. The docking domains on each terminal of the protein sequences were aligned to the HMM profile of 3 compatibility classes. The C-terminal docking domain of GZL_RS16765 and N-terminal docking domain of GZL_RS16760 were aligned to the class 2 while others were all aligned to class 1. Since Ouroboros is not predictive on class 2 now, only class 1 docking domains were analyzed. In the input MSA file, each C-terminal domain were paired with all the other N-terminal domains in the gene cluster, also integrated with the interacting dataset and the extra dataset. The results are showed in Table 3. There are three C-terminal domains were predicted to not interact with any other N-terminal domains, such as GZL_RS16785-C (C-terminal docking domain on protein GZL_RS16785), using 0.5 as the threshold. But it has the highest interaction probability with GZL_RS16780-N, which also most likely to interact with GZL_RS16785-C among all the C-terminal domains. This also happens on GZL_RS16730-C/GZL_RS16765-N and GZL_RS16760-C/GZL_RS16755-N. GZL_RS16780-C was predicted to interact with two N-terminal domains, but if it interacts with GZL_RS16750-N, there will only be three proteins in the assembly (GZL_RS16785F > GZL_RS16780 > GZL_RS16750). Considering all the possible order, if the assembly line comprises all the proteins in the gene cluster, the predicted order of protein would be GZL_RS16785F > GZL_RS16780 > GZL_RS16775 > GZL_RS16730 > GZL_RS16765 > GZL_RS16760 > GZL_RS16755 > GZL_RS16750, which is same as the order of the corresponding genes in the genome. The prediction is likely to reflect on the truth according the collinearity rules [22].

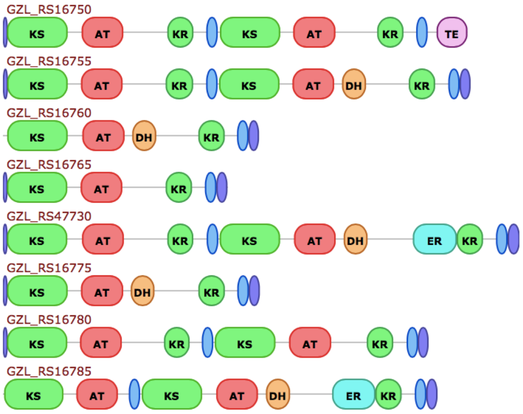


Figure 6. Gene cluster 17 in genome *Streptomyces sp*. 769 (NZ_CP003987) in antiSMASH database. Each line represents a PKS protein, and the domains on them were marked in rounded rectangles.

Table 3. Interaction probabilities of PKSs in NZ_CP003987 gene cluster 17 predicted by Ouroboros. C-terminal docking domains are in lines and N-terminals are in columns. The floats in red and orange indicates the interaction of a docking domain pair; the floats in blue indicates the highest probability of interaction of that C and N terminal docking domain.

C N	GZL_RS16780	GZL_RS16775	GZL_RS16730	GZL_RS16765	GZL_RS16760	GZL_RS16755	GZL_RS16750[L]
GZL_RS16785[F]	0.0813	1.3e-07	1.8e-05	3.2e-10	-	8.4e-08	5.5e-10
GZL_RS16780		0.9983	1.3e-06	3.2e-09	-	3.5e-06	0.9180
GZL_RS16775	3.5e-05		0.9742	6.2e-09	-	2.3e-07	9.6e-10
GZL_RS16730	9.5e-09	9.5e-09		1.6e-05	-	3.7e-07	9.7e-08
GZL_RS16765	-	-	-	-	-	-	-
GZL_RS16760	1.4e-08	6.4e-10	4.2e-06	1.4e-07	-	0.1012	4.8e-11
GZL_RS16755	2.5e-08	0.9954	8.4e-08	1.7e-07	-		0.9995

Conclusion

We aimed to adapt Ouroboros to predict PKS PPI in this study with the assumption that docking domain and KS-ACP together determine the protein interaction. Although KS-ACP pairs could not be used to predict PPI, Ouroboros showed reasonable performance on PPI prediction only based on the class 1 docking domains. We also found a residue pair that contributes to the PPI interaction specificity. Lacking class 2 docking domain sequences, the algorithm for now could only work on PKS with class 1 docking domains. We showed that the performance can be remarkably improved by integrating more sequences. Therefore, more genomes sequenced and PKS gene clusters annotated in the future could enable the prediction on other classes. When clustering the docking domains, the class 3 seemed mixed with domains belong to neither class. It is still an open question whether this is a problem of the clustering method or whether there is a "class 3". If the sequences of "class 3" cannot be properly aligned, their interaction would not likely to be predicted by Ouroboros or other correlated mutation analysis algorithm. Considering that there are still interactions between different compatibility classes of docking domains and the interacting docking domain pairs not always lead to the functional PKS PPI, it is worthwhile to investigate whether there are other domain pairs contributing to the PPI in future's study.

Acknowledgement

I would like to thank my supervisors Marnix Medema and Aalt-Jan van Dijk for their patience, insightful guidance throughout this project, and valuable feedback. It is their support that makes me more enjoy this period. I would also like to thank Miguel Correa Marrero for the useful tools and helpful advices he provided.

References

[1] McCullagh, M. (2008). Natural product pharmaceuticals-the third generation. Drug Disc. World.

[2] Fischbach, M. A., & Walsh, C. T. (2006). Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. *Chemical reviews*, 106(8), 3468-3496.

[3] Weissman, K. J. (2016). Genetic engineering of modular PKSs: from combinatorial biosynthesis to synthetic biology. *Natural product reports*, 33(2), 203-230.

[4] Yadav, G., Gokhale, R. S., & Mohanty, D. (2003). Computational approach for prediction of domain organization and substrate specificity of modular polyketide synthases. *Journal of molecular biology*, 328(2), 335-363.

[5] Menzella, H. G., & Reeves, C. D. (2007). Combinatorial biosynthesis for drug development. *Current opinion in microbiology*, 10(3), 238-245.

[6] Thattai, M., Burak, Y., & Shraiman, B. I. (2007). The origins of specificity in polyketide synthase protein interactions. *PLoS computational biology*, 3(9), e186.

[7] Yadav, G., Gokhale, R. S., & Mohanty, D. (2009). Towards prediction of metabolic products of polyketide synthases: an in silico analysis. *PLoS computational biology*, 5(4), e1000351.

[8] Rowe, C. J., Böhm, I. U., Thomas, I. P., Wilkinson, B., Rudd, B. A., Foster, G., ... & Staunton, J. (2001). Engineering a polyketide with a longer chain by insertion of an extra module into the erythromycin-producing polyketide synthase. *Chemistry & biology*, 8(5), 475-485.

[9] Thomas, I., Martin, C. J., Wilkinson, C. J., Staunton, J., & Leadlay, P. F. (2002). Skipping in a hybrid polyketide synthase: evidence for ACP-to-ACP chain transfer. *Chemistry & biology*, 9(7), 781-787.

[10] Kapur, S., Chen, A. Y., Cane, D. E., & Khosla, C. (2010). Molecular recognition between ketosynthase and acyl carrier protein domains of the 6-deoxyerythronolide B synthase. *Proceedings of the National Academy of Sciences*, 107(51), 22066-22071.

[11] Kapur, S., Lowry, B., Yuzawa, S., Kenthirapalan, S., Chen, A. Y., Cane, D. E., & Khosla, C. (2012). Reprogramming a module of the 6-deoxyerythronolide B synthase for iterative chain elongation. *Proceedings of the National Academy of Sciences*, 109(11), 4110-4115.

[12] Robbins, T., Liu, Y. C., Cane, D. E., & Khosla, C. (2016). Structure and mechanism of assembly line polyketide synthases. *Current opinion in structural biology*, 41, 10-18.

[13] Marrero, M. C., Immink, R. G., de Ridder, D., & van Dijk, A. D. (2018). Improving intermolecular contact prediction through protein-protein interaction prediction using coevolutionary analysis with expectation-maximization. *bioRxiv*, 254789.

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

[15] Blin, K., Medema, M. H., Kottmann, R., Lee, S. Y., & Weber, T. (2016). The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. *Nucleic acids research*, 2016: doi: 10.1093/nar/gkw960.

[16] Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Bruccoleri, R., ... & Breitling, R. (2015). antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic acids research*, 43(W1), W237-W243.

[17] Alekseyev, V. Y., Liu, C. W., Cane, D. E., Puglisi, J. D., & Khosla, C. (2007). Solution structure and proposed domain–domain recognition interface of an acyl carrier protein domain from a modular polyketide synthase. *Protein Science*, 16(10), 2093-2107.

[18] Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13), 1658-1659.

[19] Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics (Oxford, England)*, 14(9), 755-763.

[20] Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput *Nucleic Acids Res*. 32(5):1792-1797.

[21] Broadhurst, R. W., Nietlispach, D., Wheatcroft, M. P., Leadlay, P. F., & Weissman, K. J. (2003). The structure of docking domains in modular polyketide synthases. *Chemistry & biology*, 10(8), 723-731.

[22] Aparicio, J. F., Fouces, R., Mendes, M. V., Olivera, N., & Martín, J. F. (2000). A complex multienzyme system encoded by five polyketide synthase genes is involved in the biosynthesis of the 26-membered polyene macrolide pimaricin in Streptomyces natalensis. *Chemistry & biology*, 7(11), 895-905.

[23] Miyanaga, A., Kudo, F., & Eguchi, T. (2018). Protein–protein interactions in polyketide synthase–nonribosomal peptide synthetase hybrid assembly lines. *Natural product reports*.

[24] Scotti, C., Piatti, M., Cuzzoni, A., Perani, P., Tognoni, A., Grandi, G., ... & Albertini, A. M. (1993). A Bacillus subtilis large ORF coding for a polypeptide highly similar to polyketide synthases. *Gene*, 130(1), 65-71.

[25] Piel, J. (2010). Biosynthesis of polyketides by trans-AT polyketide synthases. *Natural product reports*, *27*(7), 996-1047.

[26] Whicher, J. R., Smaga, S. S., Hansen, D. A., Brown, W. C., Gerwick, W. H., Sherman, D. H., & Smith, J. L. (2013). Cyanobacterial polyketide synthase docking domains: a tool for engineering natural product biosynthesis. *Chemistry & biology*, 20(11), 1340-1351.

[27] Buchholz, T. J., Geders, T. W., Bartley III, F. E., Reynolds, K. A., Smith, J. L., & Sherman, D. H. (2009). Structural basis for binding specificity between subclasses of modular polyketide synthase docking domains. *ACS chemical biology*, 4(1), 41-52.

[28] Weissman, K. J. (2006). Single amino acid substitutions alter the efficiency of docking in modular polyketide biosynthesis. *ChemBioChem*, 7(9), 1334-1342.

[29] Wu, N., Cane, D. E., & Khosla, C. (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(15), 5056-5066.

[30] Dodge, G. J., Maloney, F. P., & Smith, J. L. (2018). Protein–protein interactions in “cis-AT” polyketide synthases. *Natural product reports*.

[31] Monastyrskyy, B., D'Andrea, D., Fidelis, K., Tramontano, A., & Kryshtafovych, A. (2016). New encouraging developments in contact prediction: Assessment of the CASP 11 results. *Proteins: Structure, Function, and Bioinformatics*, 84, 131-144.

[32] Dutta, S., Whicher, J. R., Hansen, D. A., Hale, W. A., Chemler, J. A., Congdon, G. R., ... & Skiniotis, G. (2014). Structure of a modular polyketide synthase. *Nature*, 510(7506), 512.

[33] Burger, L., & Van Nimwegen, E. (2008). Accurate prediction of protein–protein interactions from sequence alignments using a Bayesian method. *Molecular systems biology*, 4(1), 165.

[34] Anand, S., Prasad, M. V. R., Yadav, G., Kumar, N., Shehara, J., Ansari, M. Z., & Mohanty, D. (2010). SBSPKS: structure based sequence analysis of polyketide synthases. *Nucleic acids research*, 38(suppl_2), W487-W496.

[35] Menzella, H. G., Reid, R., Carney, J. R., Chandran, S. S., Reisinger, S. J., Patel, K. G., ... & Santi, D. V. (2005). Combinatorial polyketide biosynthesis by de novo design and rearrangement of modular polyketide synthase genes. *Nature biotechnology*, 23(9), 1171.

[36] Klaus, M., Ostrowski, M. P., Austerjost, J., Robbins, T., Lowry, B., Cane, D. E., & Khosla, C. (2016). Protein-protein interactions, not substrate recognition, dominate the turnover of chimeric assembly line polyketide synthases. *Journal of Biological Chemistry*, 291(31), *16404-16415.*

Supplements

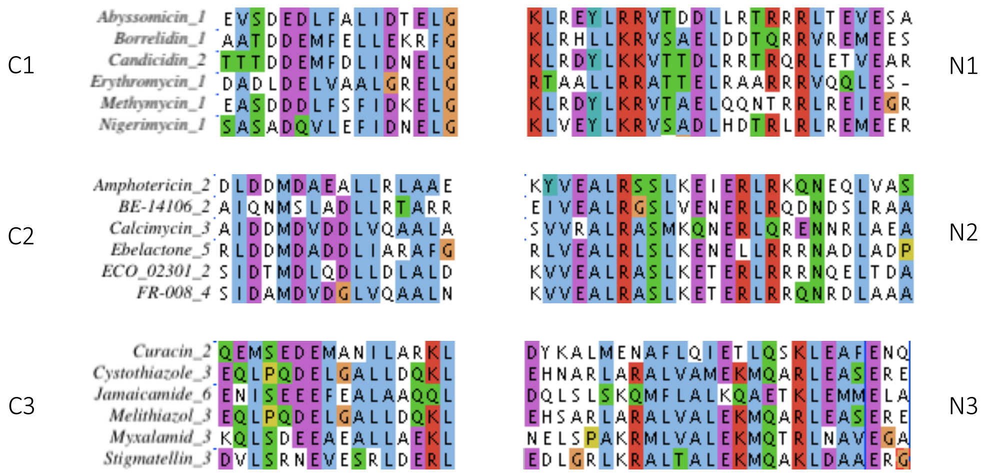
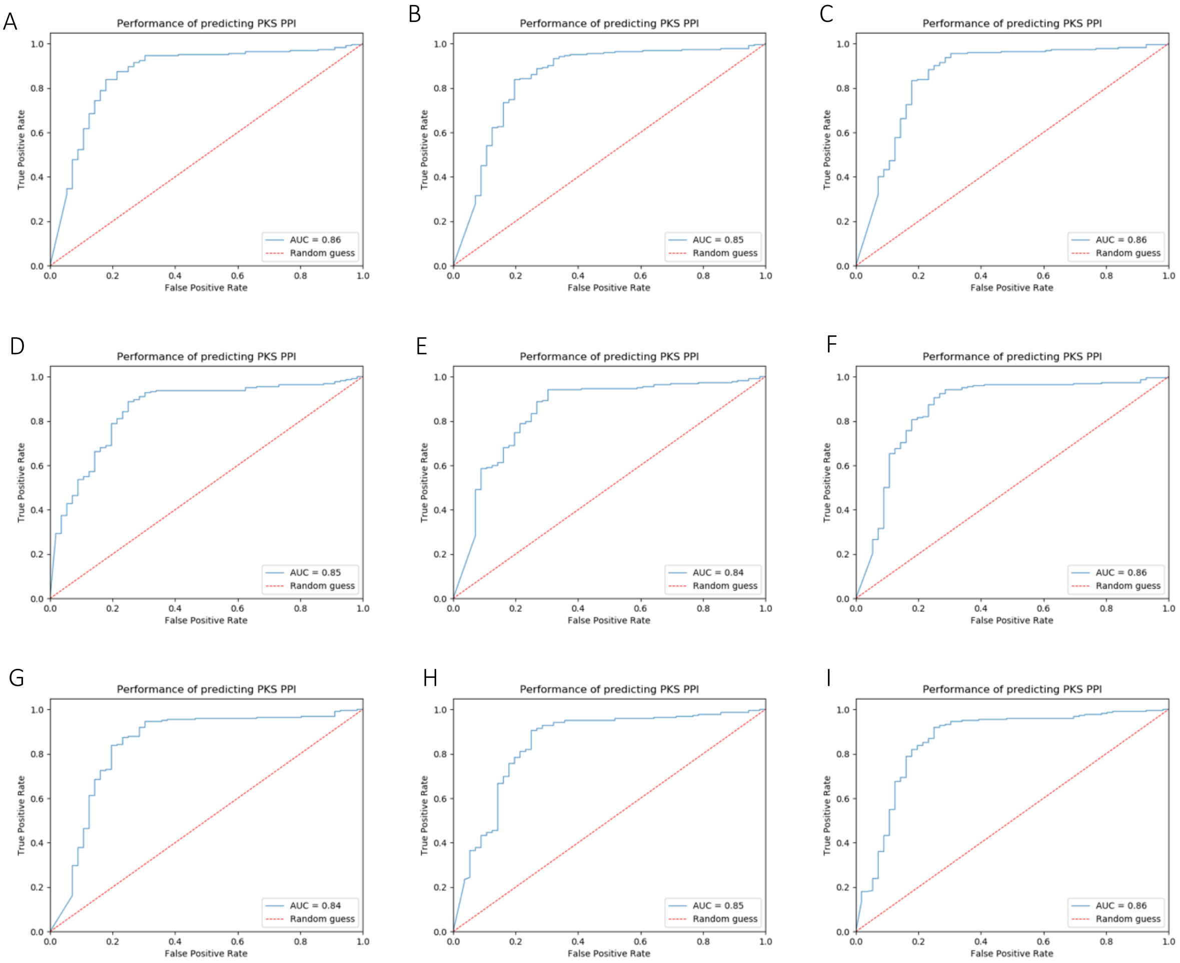

C1
Abyssomicin_1 EVSDEDLFALIDTELG KLREYLRRVTDDLLRTRRRLTEVESA
Borrelidin_1 AATDDEMFELLEKRFG KLRHLLKRVSAELDDTQRRVREMEES
Candicidin_2 TTTDDEMFDLIDNELG KLRDYLKKVTTDLRRTRQRLETVEAR
Erythromycin_1 DADLDELVAALGRELG RTAALLRRATTELRAARRRVQQLES-
Methymycin_1 EASDDDLFSFIDKELG KLRDYLKRVTAELQQNTRRLREIEGR
Nigerimycin_1 SASADQVLEFIDNELG KLVEYLKRVSADLHDTRLRLREMEER
N1
C2
Amphotericin_2 DLDDMDAEALLRLAAE KYVEALRSSLKEIERLRKQNEQLVAS
BE-14106_2 AIQNMSLADLLRTARR EIVEALRGSLVENERLRQDNDSLRAA
Calcimycin_3 AIDDMDVDDLVQAALA SVVRALRASMKQNERLQRENNRLAEA
Ebelactone_5 RLDDMDADDLIARAFG RLVEALRLSLKENELLRRRNADLADP
ECO_02301_2 SIDTMDLQDLLDLALD KVVEALRASLKETERLRRRNQELTDA
FR-008_4 SIDAMDVDGLVQAALN KVVEALRASLKETERLRRQNRDLAAA
N2
C3
Curacin_2 QEMSEDEMANILARKL DYKALMENAFLQIETLQSKLEAFENQ
Cystothiazole_3 EQLPQDELGALLDQKL EHNARLARALVAMEKMQARLEASERE
Jamaicamide_6 ENISEEEFEALAAQQL DQLSLSKQMFLALKQAETKLEMMELA
Melithiazol_3 EQLPQDELGALLDQKL EHSARLARALVALEKMQARLEASERE
Myxalamid_3 KQLSDEEAEALLAEKL NELSPAKRMLVALEKMQTRLNAVEGA
Stigmatellin_3 DVLSRNEVESRLDERL EDLGRLKRALTALEKMQAKLDAAERG
N3

Figure 1. Example of multiple sequence alignment of docking domains. Left: 16-residue C-terminal docking domains from 3 different compatible classes (C1, C2, C3). Right: 26-residue N-terminal docking domains from 3 different compatible classes (N1, N2, N3). Each line represents an interacting docking domain pair. The labels consist of the synthetic assembly lines and the PPI interface index. For example, *Abyssomicin_1* refers to the first PPI interface of Abyssomicin synthetic assembly line.



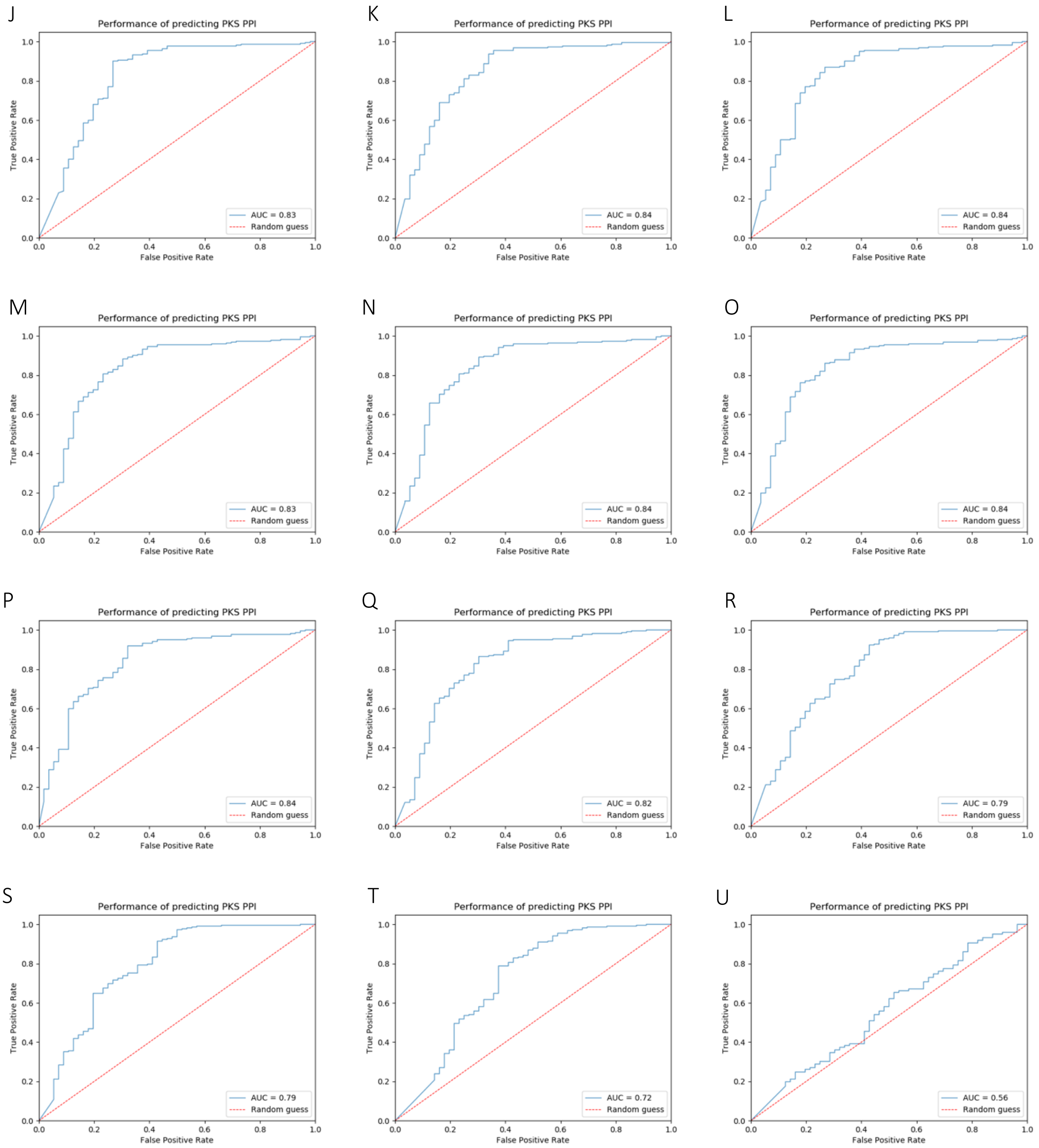


Figure 2. Predictive performance with the absence of different contact pairs on docking domains. From A to U, one more residue pair was removed from the MSA each prediction. The order of residue pairs removed was {14,14}, {1,7}, {14,18}, {14,21}, {5,1}, {8,2}, {8,1}, {8,5}, {8,6}, {14,13}, {10,14}, {7,11}, {15,12}, {7,7}, {15,13}, {10,10}, {15,4}, {11,9}, {11,13}, {12,5}, {4,3}. For example, A is the performance with the absence of {14,14}, B is the performance with the absence of {14,14} and {1,7}, and C is the performance with the absence of {14,14}, {1,7}, and {14,18}. The last figure U shows performance with the absence of all the residue pairs mentioned above.

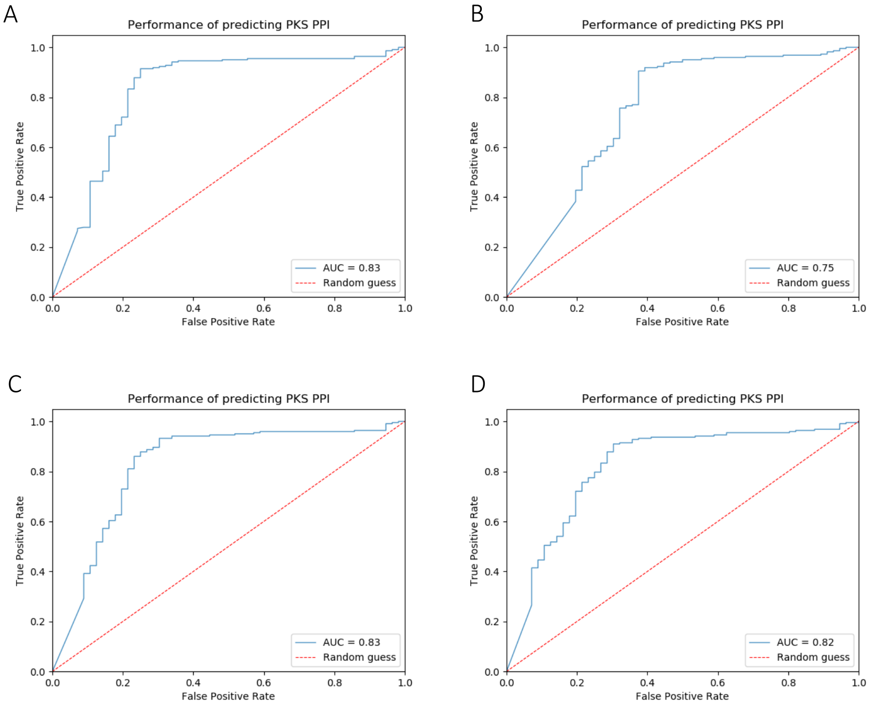


Figure 3. Predictive performance of PPI with the absence of residue pair A) {11,9}, B) {12,5}, C) {5,1}, D) {14,14} on docking domains.