

Figure S2: Figure depicting extraction of all putative lanthionine cross-link forming substrings of the type Ser/Thr-(X)_n-Cys or Cys-(X)_n-Ser/Thr from the core peptide sequence of the lanthipeptide nisin A. Arrows in blue color indicate Ser/Thr and Cys residues on the unmodified core peptide which form lanthionine cross-links in the structure of nisin A, while arrows in red color indicate sites which are not crosslinked in nisin A. All sub-sequences of the type Ser/Thr-(X)_n-Cys or Cys-(X)_n-Ser/Thr have been listed. The sub-sequences listed in blue correspond to correct cross-links, while those listed in red are incorrect cross-links (not found in nisin A).

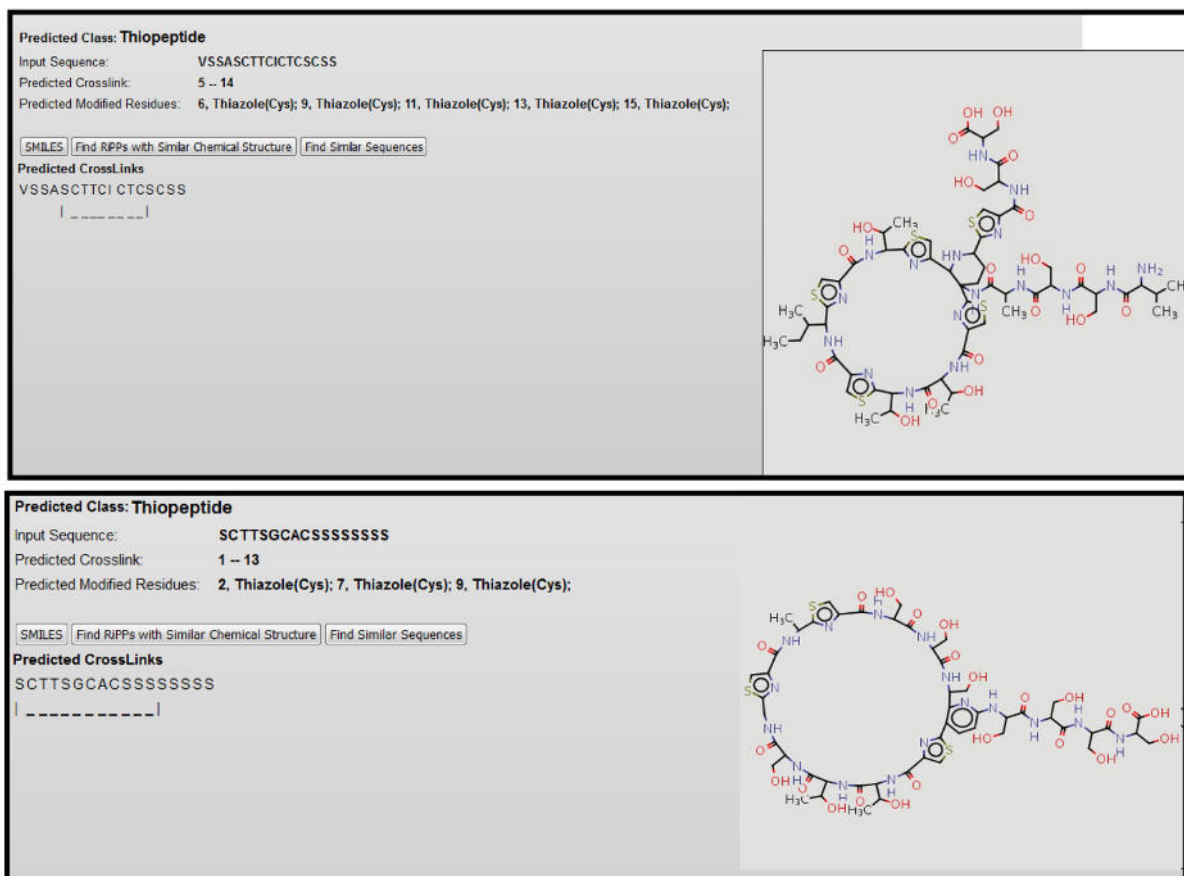


Figure S3: Screenshots depicting cross-link predictions results from RiPPMiner for two thiopeptides. The core peptide sequences of Siomycin A and A10255B have been used as input.

Predicted Class: Lasso peptide
SEQUENCE: MHTPIISETVQPKTAGLIVLGKASAETRGLSQGVPEPDIGQTYFEESRINQD
[Find Similar Sequences](#)

MODEL 1
Cleavage Site: 28
Leader Peptide: MHTPIISETVQPKTAGLIVLGKASAETR
Core Peptide: GLSQGVPEPDIGQTYFEESRINQD
Predicted Crosslinks: [SMILES](#)
[Find RiPPs with Similar Chemical Structure](#)
GLSQGVPEPDI GQTYFEESRI NQD
| _ _ _ _ _ |

MODEL 2
Cleavage Site: 28
Leader Peptide: MHTPIISETVQPKTAGLIVLGKASAETR
Core Peptide: GLSQGVPEPDIGQTYFEESRINQD
Predicted Crosslinks: [SMILES](#)
[Find RiPPs with Similar Chemical Structure](#)
GLSQGVPEPDI GQTYFEESRI NQD
| _ _ _ _ _ |

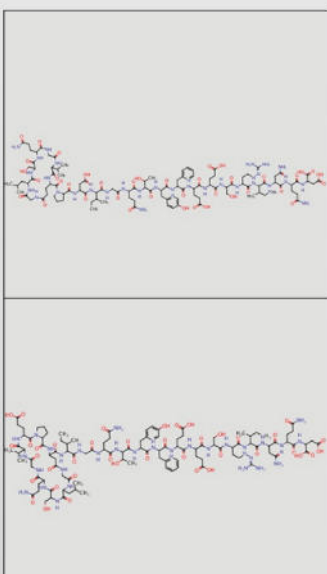


Figure S4: Results from RiPPMiner webserver for RiPP class, cleavage and cross-link prediction for lasso peptide. The precursor peptide sequences of Astexin has been used as input.

Predicted Class: Cyanobactin

[Find Similar Sequences](#)

CORE 1

Sequence: MDKKNILPHQGKPVLRRTTNGKLPShLAELSEEALGGNGVDASACMPCYPSYDGVDS**VCMP**CYPSYDGVDSVCMPCYPSYDAAE

Core Peptide: **VCMP**CYP

Core Position: 56-64

Modified Residues: 2,thiazol(In)e(Cys) 5,thiazol(In)e(Cys)

Predicted Crosslinks: [SMILES](#)

[Find RiPPs with Similar Chemical Structure](#)

VCMP CYP
| _ _ _ _ |

CORE 2

Sequence: MDKKNILPHQGKPVLRRTTNGKLPShLAELSEEALGGNGVDAS**ACMP**CYPSYDGVDSVCMPCYPSYDGVDSVCMPCYPSYDAAE

Core Peptide: **ACMP**CYP

Core Position: 43-49

Modified Residues: 2,thiazol(In)e(Cys) 5,thiazol(In)e(Cys)

Predicted Crosslinks: [SMILES](#)

[Find RiPPs with similar Chemical Structure](#)

ACMP CYP
| _ _ _ _ |

CORE 3

Sequence: MDKKNILPHQGKPVLRRTTNGKLPShLAELSEEALGGNGVDASACMPCYPSYDGVDSVCMPCYPSYDGVDS**VCMP**CYPSYDAAE

Core Peptide: **VCMP**CYP

Core Position: 73-79

Modified Residues: 2,thiazol(In)e(Cys) 5,thiazol(In)e(Cys)

Predicted Crosslinks: [SMILES](#)

[Find RiPPs with Similar Chemical Structure](#)

VCMP CYP
| _ _ _ _ |

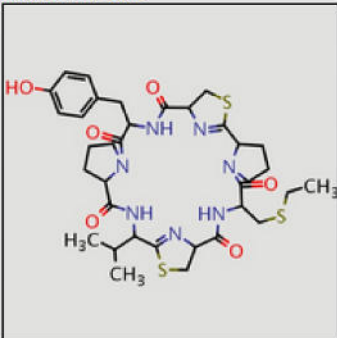
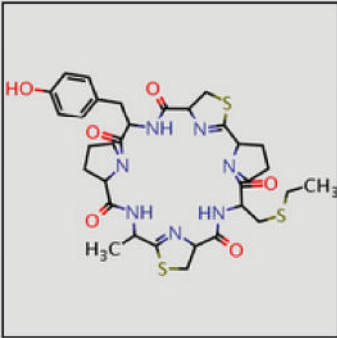
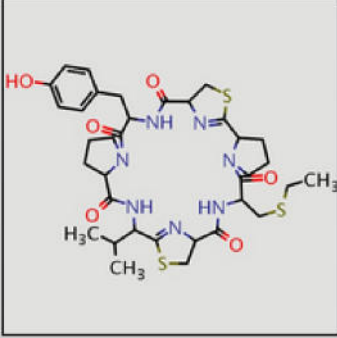




Figure S5: Results from RiPPMiner webserver for RiPP class, cleavage and cross-link prediction for cyanobactin. The precursor peptide sequences of Aesturamide has been used as input.

RiPP PRISM

Cluster 1
 Class: lanthipeptide
 Cluster size: 24 proteins
 From species: 602776294-001-001
 Source: host genome (Lipari)

Bioinformatic assembly:
 Genes: 24
 ORFs: 24
 Clusters: 1

Genomic:
 [Genomic map showing gene locations]

Legend:

- Red: Homologous peptide
- Blue: Peptide
- Green: Peptide
- Orange: Peptide
- Yellow: Peptide
- Light blue: Peptide
- Light green: Peptide
- Light orange: Peptide
- Light yellow: Peptide
- Light purple: Peptide
- Light red: Peptide
- Light blue: Peptide
- Light green: Peptide
- Light orange: Peptide
- Light yellow: Peptide
- Light purple: Peptide
- Light red: Peptide

Predicted cluster product:
 Compound and Cluster Key: T32
 SMILES: [SMILES string]

Download:
 Download cluster product
 Download cluster key
 Download SMILES

RiPPMiner

A Bioinformatics Resource for Deciphering Chemical Structures of RiPPs

Keyword Search: [input field] [GO]

HOME TOOLS DATABASE BENCHMARK DOWNLOAD

Predicted Class: lanthipeptideA

Sequence: MSTKDFILDUVYKSKDSDGASPRHTSBLLTPOKTDALHGGHWKDTCHSHVSK

Chemical info (IEMer): S-CASPRH-SBL

Leader peptide: MSTKDFILDUVYKSKDSDGASPR

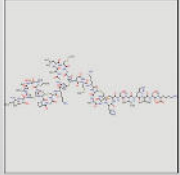
Core peptide: HSBLLTPOKTDALHGGHWKDTCHSHVSK

Predicted CrossLinks:

Find Similar Sequences
 Find similar sequences with similar chemical structure

IT SI SLGT PGKHTGALMGCHMKTATGHC SI HV SK

CrossLinks:
 CrossLink: 3-7 Ser-Ops
 CrossLink: 8-11 Thr-Cys
 CrossLink: 13-19 Thr-Cys
 CrossLink: 23-26 Thr-Cys
 CrossLink: 25-28 Thr-Cys



antiSMASH

Cluster 1: Lanthipeptide

Genomic map: [Genomic map showing gene locations]

Legend:

- Red: Homologous peptide
- Blue: Peptide
- Green: Peptide
- Orange: Peptide
- Yellow: Peptide
- Light blue: Peptide
- Light green: Peptide
- Light orange: Peptide
- Light yellow: Peptide
- Light purple: Peptide
- Light red: Peptide
- Light blue: Peptide
- Light green: Peptide
- Light orange: Peptide
- Light yellow: Peptide
- Light purple: Peptide
- Light red: Peptide

antiSMASH Results:

Cluster 1: Lanthipeptide

Genomic map: [Genomic map showing gene locations]

Legend:

- Red: Homologous peptide
- Blue: Peptide
- Green: Peptide
- Orange: Peptide
- Yellow: Peptide
- Light blue: Peptide
- Light green: Peptide
- Light orange: Peptide
- Light yellow: Peptide
- Light purple: Peptide
- Light red: Peptide
- Light blue: Peptide
- Light green: Peptide
- Light orange: Peptide
- Light yellow: Peptide
- Light purple: Peptide
- Light red: Peptide

antiSMASH Results:

Cluster 1: Lanthipeptide

Genomic map: [Genomic map showing gene locations]

Legend:

- Red: Homologous peptide
- Blue: Peptide
- Green: Peptide
- Orange: Peptide
- Yellow: Peptide
- Light blue: Peptide
- Light green: Peptide
- Light orange: Peptide
- Light yellow: Peptide
- Light purple: Peptide
- Light red: Peptide
- Light blue: Peptide
- Light green: Peptide
- Light orange: Peptide
- Light yellow: Peptide
- Light purple: Peptide
- Light red: Peptide

Figure S6: Screenshots depicting results for lanthipeptide cleavage and cross-link predictions by RiPP-PRISM, antiSMASH and RiPPMiner. The precursor peptide sequence of nisin has been used as input.

SUPPLEMENTARY METHODS

RiPP Identification

To distinguish RiPP from other proteins and peptides SVM model was trained using amino acid composition and dipeptide frequencies as feature vectors. The model was trained on 147 precursor proteins of RiPPs (positive set) and 4070 non-RiPP peptides and proteins (negative set). The proteins in negative set were chosen so that they are similar to RiPP in length (10 to 100 amino acids), are manually curated (Swiss-Prot entries) and belong to families other than RiPP like 30s ribosomal proteins, matrix proteins, cytochrome B, ATP synthase subunit and acyl carrier protein. SVM Model was trained by including the ‘cost factor’ to minimize the effect of differences in numbers in positive and negative set. **Supplementary Table S1** shows the benchmarking results for RiPP identification using two fold cross validation approach on a test set containing 146 RiPPs and 4070 non-RiPP polypeptides. As can be seen below, even though sensitivity and specificity values were high, precision and MCC were low because the negative dataset was much larger than the positive dataset.

Supplementary Table S1: Results of two fold cross validation for identification of RiPPs

		SVM model trained on Set 1		
	Total	In test set (Set2)	True positive	False negative
Positive dataset	293	146	137	9
		In test set (Set2)	False positive	True negative
30s_ribosomal_protein	2763	1346	8	1338
40s_ribosomal_protein	87	40	3	37
50s_ribosomal_protein	4122	2116	27	2089
acyl_carrier_protein	282	139	69	70
Amylin	12	4	1	3
ATP_synthase_subunit	374	192	148	44
Calcitonin	24	17	17	0
cytochrome_b_protein	458	208	131	77
matrix_protein	18	8	2	6
Total	8140	4070	406	3664
Sensitivity			0.938356164	
Specificity			0.9002457	
Precision			0.252302026	
MCC			0.464453191	
AUC			0.967	

In order to deal with the imbalanced positive and negative datasets, the predictions were carried out by randomly dividing the negative dataset into 27 different sets such that each time negative dataset was comparable in size to the positive set. For each set two-fold cross validation was performed. **Supplementary Table S2** shows results for each of the 27 predictions. This gave the average AUC of 0.96 and average precision and MCC values were also above 0.8.

Supplementary Table S2: Results of two fold cross validation for RiPP identification when negative dataset was randomly divided into 27 equal sets.

	Negative Total	Positive Total	True positive	False negative	True Negative	False Positive	Sensitivity	Specificity	Precision	MCC	AUC
set1	151	146	137	9	134	17	0.94	0.89	0.89	0.84	0.97
set2	151	146	137	9	136	15	0.94	0.90	0.90	0.85	0.97
set3	151	146	137	9	139	12	0.94	0.92	0.92	0.87	0.97
set4	151	146	137	9	144	7	0.94	0.95	0.95	0.90	0.98
set5	151	146	137	9	140	11	0.94	0.93	0.93	0.87	0.98
set6	151	146	137	9	135	16	0.94	0.89	0.90	0.85	0.96
set7	151	146	137	9	138	13	0.94	0.91	0.91	0.86	0.97
set8	151	146	137	9	138	13	0.94	0.91	0.91	0.86	0.97
set9	151	146	137	9	136	15	0.94	0.90	0.90	0.85	0.97
set10	151	146	137	9	142	9	0.94	0.94	0.94	0.89	0.98
set11	151	146	137	9	139	12	0.94	0.92	0.92	0.87	0.97
set12	151	146	137	9	138	13	0.94	0.91	0.91	0.86	0.97
set13	151	146	137	9	136	15	0.94	0.90	0.90	0.85	0.96
set14	151	146	137	9	131	20	0.94	0.87	0.87	0.82	0.96
set15	151	146	137	9	134	17	0.94	0.89	0.89	0.84	0.96
set16	151	146	137	9	133	18	0.94	0.88	0.88	0.83	0.96
set17	151	146	137	9	121	30	0.94	0.80	0.82	0.77	0.95
set18	151	146	137	9	139	12	0.94	0.92	0.92	0.87	0.97
set19	151	146	137	9	131	20	0.94	0.87	0.87	0.82	0.95
set20	151	146	137	9	138	13	0.94	0.91	0.91	0.86	0.98
set21	150	146	137	9	139	11	0.94	0.93	0.93	0.87	0.97
set22	150	146	137	9	138	12	0.94	0.92	0.92	0.87	0.98
set23	150	146	137	9	130	20	0.94	0.87	0.87	0.82	0.95
set24	150	146	137	9	135	15	0.94	0.90	0.90	0.85	0.97
set25	150	146	137	9	134	16	0.94	0.89	0.90	0.85	0.97
set26	150	146	137	9	131	19	0.94	0.87	0.88	0.83	0.96
set27	150	146	137	9	135	15	0.94	0.90	0.90	0.85	0.97
Average							0.94	0.90	0.90	0.85	0.97

RiPP Class Prediction

Multiclass SVM Classifier model was generated using SVM^{multiclass}. **Supplementary Table S3** shows the results of benchmarking for RiPP class prediction.

Supplementary Table S3: Benchmarking for RiPP class prediction using Leave-One-Out

	Class	TP	FP	TN	FN
1	Lanthipeptide B	56	5	170	7
2	Lanthipeptide A	30	0	208	0
3	Lanthipeptide C	8	1	225	4
4	Linardin	10	3	220	5
5	Cyanobactin	42	24	172	0
6	Sactipeptide	0	0	233	5
7	Microcin	6	1	231	0
8	Lasso peptide	16	0	207	15
9	Bacterial_head_to_tail_cyclized	12	1	225	0
10	Auto_inducing_peptide	3	0	234	1
11	ComX	4	1	233	0
12	Thiopeptide	13	1	224	0

Lanthipeptide Cleavage Prediction

Leader cleavage site is defined by 12mer motifs with 6 residues upstream and downstream to the actual cleavage site. To train the SVM total 115 lanthipeptide precursor peptides were used to generate all possible 12mers. This resulted in 103 true unique 12mer cleavage sites (positive set) and 4524 data in negative set. SVM Model was trained by including the 'cost factor' to minimize the effect of differences in numbers in positive and negative set. To test the model, almost equal number of positive and negative cleavage motifs (12mers) was divided into 43 sets and for each set two-fold cross validation was performed. This gave the average AUC of **0.93**.

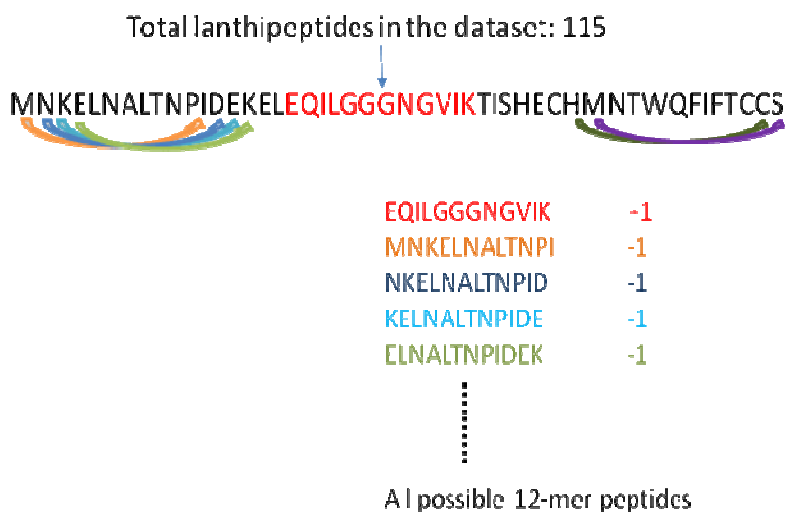


Figure. Generation of positive and negative 12mers from input data and training the SVM.

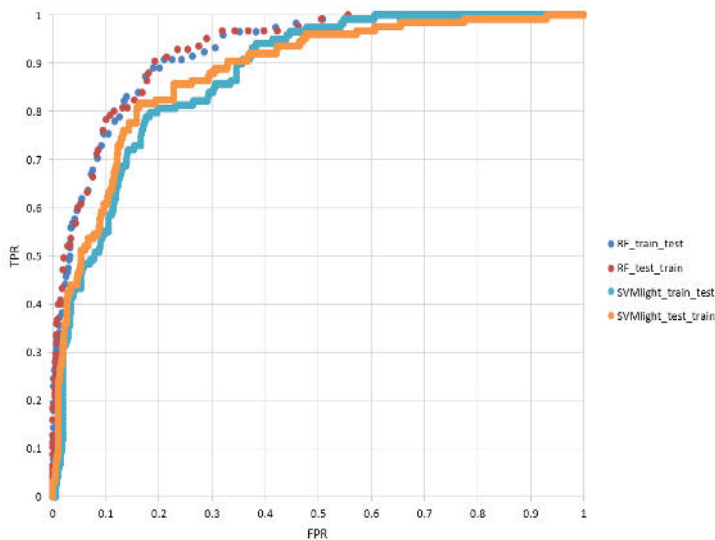
Further Lanthipeptide wise Leave-one-out was performed where 1 lanthipeptide was used for testing and remaining were used in model training. Hence training set will have all possible 12mers from 114 Lanthipeptides whereas test set will contain 12mers from 1 lanthipeptide. Actual cleavage site was scored maximum in around 75.7% cases (total 87) while 85.2% (total 98) cleavage sites were present in top 2 maximum scoring 12mers.

Lanthipeptide Cross-link Prediction

All possible Ser/Thr-(X)_n-Cys and Ser/Thr-(X)_n-Cys fragments were generated for 93 lanthipeptides which resulted in a total of 1576 unique fragments, of which 218 were positive and 1358 were negative fragments. A 'cost factor' was used to minimize the effect of differences in size of positive and negative dataset. The feature vectors were created using fragment-length-normalized frequencies of di-amino and mono-amino acid compositions within the fragments. These feature vectors were used to construct Random forest and SVM models. For Random Forest Model, the AUC score in leave-one-out analysis was **0.90** and the average

AUC for 2-fold cross validation was **0.78**. Similarly for SVM Model respective AUC values were **0.82** and **0.72**.

In the next analysis, 93 lanthipeptides were divided into two sets; training set containing 48 Lanthipeptides (125 positive and 751 negative fragments) and test set containing 45 (118 positive and 724 negative fragments) lanthipeptides. ROC curves and AUC scores are shown below.



Set	Positive Fragments	Negative Fragments	Total
Train	125	751	876
Test	118	724	842

Experiment	ROC
RF – train test	0.922
RF – test train	0.927
SVM light train test	0.873
SVM light test train	0.879

Cyanobactins, due to their biologically relevant activities, are considered to be one of the most promising sources of new prospective drugs. Therefore, identification and structure prediction of cyanobactins is an area of growing interest. Precursor peptide of cyanobactin might contain up to four hypervariable core sequences. Each core sequence is flanked by an N-terminal recognition sequence (RSII) and a C-terminal recognition sequence (RSIII) (Sardar et al (2015) ACS Synth Biol 4:167-76). Serine, threonine and cysteine amino acids in the core peptide are heterocyclized by heterocyclases based on the presence of recognition sequence I (RSI). To predict heterocyclized residues 28 characterized cyanobactins were used. Of the 28 peptides 21 contained heterocycles whereas 7 had no heterocycles present in them. SVM model using amino acid composition and dipeptide composition as feature was used to predict presence of heterocycles. To predict the location of RSII each sequence was fragmented into 5 mer peptides. 5 mer peptides containing RSII motif were designated as positive and the rest as negatives. From the 28 cyanobactins 52 positive and 716 negative peptides were found. 5 mer positional matrix (5*20) was used as feature vector to train the SVM model. A very similar approach was followed to predict the location of RSIII, the only difference being 4 mer positional matrix was used as feature vector to train the SVM. The models were cross validated using leave-one out method

and the AUC for detection of heterocyclization, RSII and RSIII were 1, 0.9591 and 0.9466 respectively. Once the locations of RSII and RSIII motif were determined, the sequence between these two RS was predicted as core peptide. Cyanobactin undergoes head-to-tail cyclization hence the predicted core sequence was used to predict the C-N macrocyclization. The combined predictor was then cross validated using LOO method. Of 52 core peptides from 28 characterized cyanobactin precursor genes, structure of 45 core peptides (86.54%) were predicted correctly

Lasso peptides constitute a class of RiPP whose knot like fold confers them with exceptional structural stability and interesting bioactivities. The core peptide consists of 7-9 membered macrolactam ring through which the C-terminal tail is threaded. The macrolactam ring is formed between N-terminal amino acid of the core peptide and side chain of aspartate or glutamate residues. SVM based model using 13-mer positional matrix (13*20) as feature vector was built to predict the peptide cleavage. Cross validation using leave-one out method gave an AUC value of 0.998. Examination of 31 characterized lasso peptide structures helped us in devising a rule to predict macrolactam ring formation. We used the first residue of the core peptide and side chain of aspartate or glutamate residue at 7th, 8th or 9th position to predict the ring formation. First occurrence of acidic amino acid was used in cross-link formation when more than two were present at 7th, 8th or 9th position. The combined predictor was used in cross validation using LOO method and was shown to predict the structure of 30 out of 31 (96.77%) lasso peptides correctly. Hegemann et al., (Hegemann JD et al., Biopolymers. 2013, doi: 10.1002/bip.22326) had identified and predicted the structure of 87 putative lasso peptides from proteobacterial strains. Of the identified lasso peptides 60 putative lasso peptides were not present in our database of characterized lasso peptides. For 50 (83.33%) lasso peptides our predictions matched with the predictions of Hegemann *et al.*

Thiopeptide Cross-links Prediction

In case of thiopeptides the cross-links have been predicted based on occurrence of **SC**-(X)_n-**CSC** or **SC**-(X)_n-[C/S]**SSSS**, where Ser residues marked in bold are post-translationally modified to Dha and are then crosslinked via formation of nitrogen containing six membered rings. This motif based method for thiopeptides' cross-links prediction was tested on 35 distinct thiopeptides. Out of 35, True cross-links were predicted in 28 cases thus giving the accuracy of **80%**.

Supplementary Table S4: Summary of results for prediction of cross-links in lanthipeptides.**

Sl. No.	Name	TP	FP	TN	FN
1.	Ancovenin	0	2	7	3
2.	Avermipeptin	2	0	9	0
3.	Bovicin HJ50	1	2	20	1
4.	Catenulipeptin	2	0	8	0
5.	Cinnamycin	1	2	5	2
6.	Curvopeptin	2	0	10	0
7.	Duramycin B	0	2	5	3
8.	Duramycin C	0	2	7	3
9.	Duramycin	1	2	5	2
10.	Epilancin 15X	3	0	20	0
11.	Epilancin K7	2	1	19	1
12.	Gardimycin(actagardine)	4	0	15	0
13.	Haloduracin alpha	2	0	14	1
14.	Haloduracin beta	4	0	27	0
15.	Lacticin 3147 A1	2	1	21	2
16.	Lacticin 3147 A2	2	1	27	1
17.	Lacticin 481	1	1	12	2
18.	Lactocin S	0	1	13	2
19.	Lichenicidin A1	3	0	25	1
20.	Lichenicidin A2	4	0	54	0
21.	LichenicidinVK21A1	3	0	25	1
22.	LichenicidinVK21A2	4	0	54	0
23.	Mersacidin	1	1	13	3
24.	Michiganin A	2	2	18	1
25.	Mutacin 2	1	1	7	2
26.	Mutacin B Ny266	4	0	19	0
27.	Mutacin I	4	0	22	0
28.	NAI 107	3	0	32	2
29.	NAI 112	2	0	16	0
30.	Nisin Q	5	0	37	0
31.	Nisin U	5	0	32	0
32.	Nisin Z	5	0	37	0
33.	Nukacin A	2	1	12	1
34.	Paenibacillin	5	0	36	0
35.	paenicidin B	6	0	45	0
36.	planosporicin	4	0	32	1
37.	Plantaricin W beta	2	2	17	1
38.	Ruminococcin A	2	0	8	1
39.	Salivaricin A	1	1	8	2
40.	Sap B	2	0	13	0
41.	SAP T	4	0	10	0
42.	Stackepeptin C	2	1	27	1
43.	Stackepeptin D	2	1	27	1
44.	Streptococcin A FF22	1	2	8	2
45.	Streptococcin A M49	1	2	8	2
	TOTAL	109	31	886	45

**Comparison of the predicted cross links with the cross-links in the actual structures can be viewed at http://www.nii.ac.in/~priyesh/lantipepDB/xlink_train_testRF/traintest_list_new.php