Use of pattern recognition methods for fungal adenylation domain substrate specificity predictions

Sagar Gore, Ekaterina Shelest

Systems Biology/Bioinformatics- Leibniz Institute for Natural Product Research and Infection Biology- Hans Knöll Institute (HKI), Jena, Germany.

1. Non ribosomal peptide synthetases (NRPS)

C-A-T-C-A-M-T-C-A-M-T-C-A-M-T-C-A-M-T-C-A-T-C-A-M-T-C-A-M-T-C-A-T-C-A-T-C

NRPS : megasynthetases for non ribosomal peptide (NRP) biosynthesis composed of one/many modules and each module might have following domains. Core: A, C, T, Accessory: M

A:Adenylation

Substrate selection and it's activation by adenylation

C:Condensation

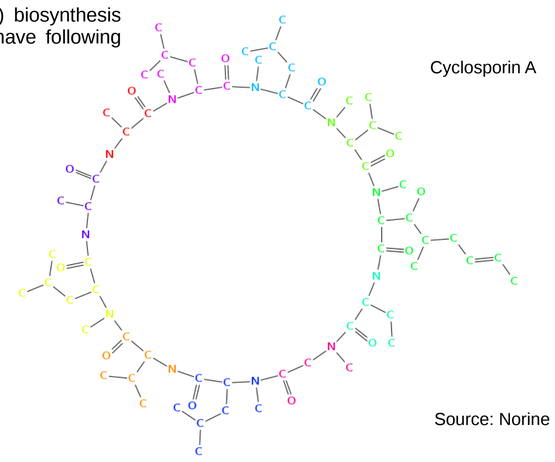
Peptide bond formation between growing peptide and monomer activated by downstream A domain

T:Thiolation

Substrate shuttling among active sites

M:Methyltransferase

Accessory domain responsible for methylation of substrates



Cyclosporin A

Source: Norine

Fig. 1 Cyclosporin A (Immunosuppressant and antibiotic activity). This compound is produced by a fungus *Tolypocladium inflatum* and used in organ transplants to prevent rejection.

2. A domain substrate specificity

NRPS code (similar to triplet codon in ribosomal peptide synthesis) was defined from bacterial PheA (shown in fig 1).

Genome mining studies: suggested many orphan biosynthetic gene clusters (BGCs) possibly encoding NRPs in bacterial/fungal genomes.

Bioinformatics tools: A domain substrate specificity (listed in Table1). These tools work **well for bacterial sequences** but **poorly for fungal sequences**.

Our goals of this study are **to decipher fungal A domain substrate specificity** and at the last to predict complete chemical structure of secondary metabolites encoded by BGCs.

PDB: 1AMU

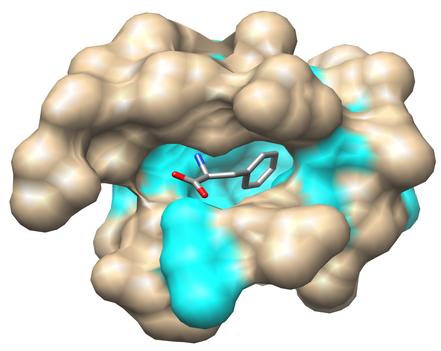


Fig. 2 Phenylalanine activating domain of gramicidin synthetase PheA (*Brevibacillus brevis*) PDB code 1AMU. 8Å residues: surface representation (tan color). 10 residues defining **NRPS code:** cyan colored patch.

3. A domain substrate specificity prediction tools

Tool	Algorithm	Dataset	Results
SANDPUMA (2017)	Decision tree	928 SQ (90 Fungal) 104 SB	Accuracy 0.84
Virtual Screening (2015)	Ligand docking	10 Crystal structures 12 Homology models 161 Ligands	Accuracy Crystal structures 1.0 Models 0.61
SEQL-NRPS (2015)	Sequence learner	537 SQ 37 SB	Accuracy 0.71
LSI (2014)	Latent semantic indexing	397 SQ 37 SB	Bacterial 0.89 Fungal 0.85
NRPSpredictor2 (2011)	Support vector machines	576 SQ (Labeled) 5096 SQ (Unlabeled) 75 SB	Bacterial F 0.94 Fungal F 0.84

Table1. NRPS A domain substrate specificity prediction tools. SB = Substrates SQ= Sequences

Introduction

4. Are fungal/bacterial binding sites similar?

To answer the above question, binding sites (34 residues) of 546 A domain sequences from NRPSpredictor2 dataset [3] were encoded by aaindex and clustered using 10K bootstrap cycles with pvclust [4] package in R.

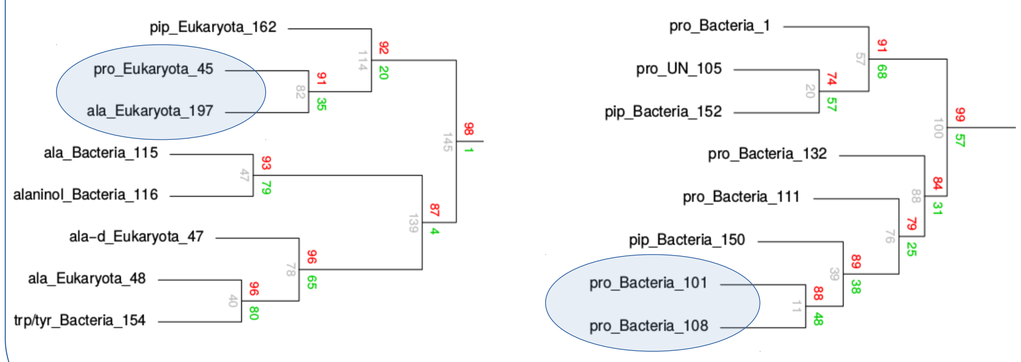
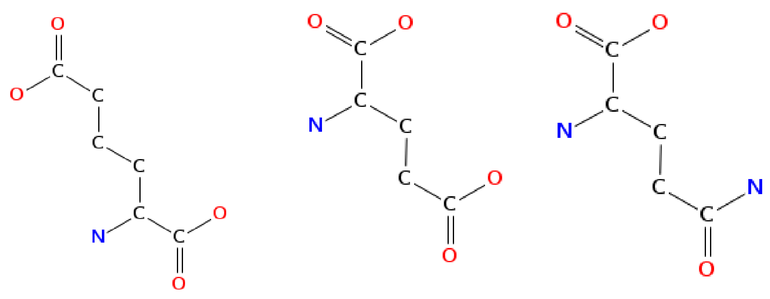


Fig. 3 Part of A domain binding site residues clustering dendrogram. Left panel (Eukaryotic) and Right panel (Bacterial). Pro/Pip substrates binding sites from bacterial and fungal sequences do not cluster together.

5. A domain substrate and NRPS code similarity

533 NRPS A domain substrates (obtained from 876 Curated and 310 Putative NRPs [1]) were transformed into SMILES chemical structure format and were encoded into Morgen fingerprints [2].

2-amino adipic acid (aad) Glutamic acid (glu) Glutamine (gln)



```
DPRHFVMRA aad Eukaryota
EPRNIVEFV aad Eukaryota
EPRHIVEFV aad Eukaryota
DPRHFVMRS aad Eukaryota
EPRNVVEFV aad Eukaryota
EPRNLVEFV aad Bacteria

DPMMWMAIN glu Bacteria
DAWHFGSVE glu Bacteria
DAWHFGGVD glu Bacteria
DAKDIGVVD glu Bacteria
DAKDLGVVD glu Bacteria
DPRHSGVVG glu Bacteria
DVWHFGRIN glu Bacteria
DLVKVASVN glu Bacteria

DGGMVGGNY gln Eukaryota
DAWQFGLID gln Bacteria
DAQDLGVVD gln Bacteria
```

Fig. 4 Three structurally similar substrates clustered together (left side) in a dendrogram and their corresponding NRPS codes (right side). 1st and 2nd **most abundant** residues in a each NRPS code column are colored Red and Blue respectively.

Results

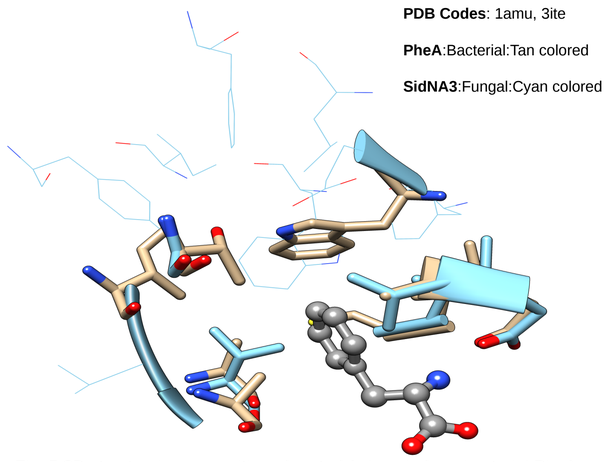
6. Inductive support vector machine (SVM)

Binding site residues (10 residue NRPS code or 34 residues within 8Å) are obtained by aligning query sequences with PheA-bacterial A domain shown in fig5.

10 residues from PheA that are used to define NRPS code are shown in tan colored sticks.

9 NRPS code residues were encoded with physicochemical properties with wold encoding
z1: Hydrophobicity z2: Size z3:Electronic properties

27 (9 residues * 3 properties) features vector was used in inductive SVM classification. There were 9 substrate classes and LOO cross validation results are shown in fig6.



PDB Codes: 1amu, 3ite

PheA:Bacterial:Tan colored

SidNA3:Fungal:Cyan colored

Fig. 5 3D structural superposition of part of A domain from bacterial PheA (tan) and fungal SidNA3 (cyan). Extra residues in SidNA3 (blue sticks). Phe substrate (grey).

7. LOO cross validation

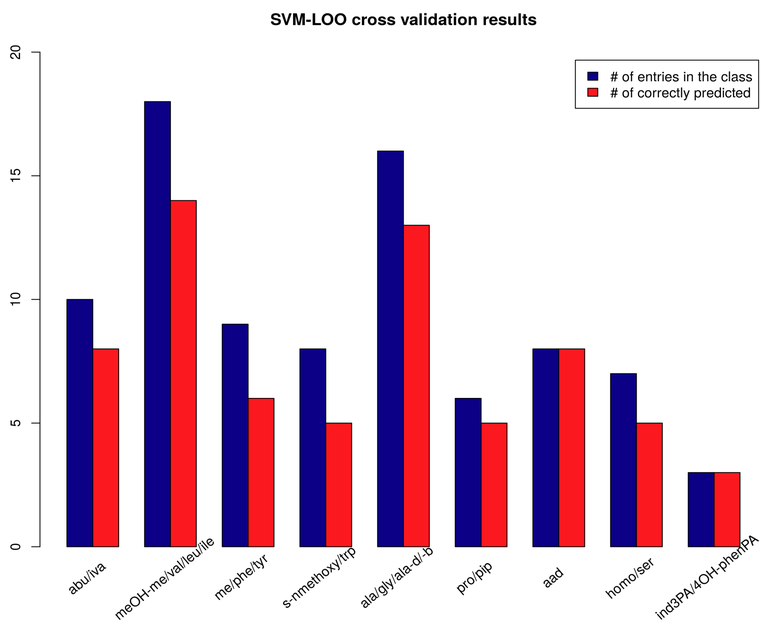


Fig. 6 Total number of entries in each class are shown and correctly predicted entries (with optimized parameters) were counted. Grid search was used to optimize the parameter C (cost penalty) while J (cost factor) was fixed. Parameter C is a trade off between maximum margin and classification error.

8. Conclusions and future prospects

i. Substrates with differential interactions in bacterial and fungal binding sites (BS) are shown below.

Similar BS	Aad	Val	Trp	Phe	-	Cys	Leu
Different BS	Gln	Ala	Tyr	Pro	Ser		

Table2: List of substrates for which binding sites are similar or/and different in bacterial and fungal sequences.

ii. Though prediction accuracy is good for some substrates (e.g aad) to further improve an accuracy for other substrates phylogeny information and more labeled/unlabeled data would be helpful in semi supervised learning approaches.

References

[1] Caboche, S., et al. "NORINE: a database of nonribosomal peptides." (2008)

[2] Rogers, D. et. al. "Extended-Connectivity Fingerprints." (2010)

[3] Röttig, M. et al. "NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity."(2011)

[4] Suzuki R. et al. " "Pvclust: An R Package for Hierarchical Clustering with p Values Via Multiscale Bootstrap Resampling, R Package Version 1.2-1." (2006)

Acknowledgements

This work was supported by the Collaborative Research Center ChemBioSys (CRC 1127) funded by Deutche Forschung Gemeinschaft (DFG). This work was carried out at Leibniz Institute for Natural Product Research and Infection Biology- Hans Knöll Institute (HKI), Jena, Germany.

Contact

sagar.gore@leibniz-hki.de
ekaterina.shelest@leibniz-hki.de

Leibniz Institute for Natural Product Research and Infection Biology
Hans Knöll Institute