

Pattern recognition methods for the prediction of chemical structures of fungal secondary metabolites

Dissertation

To Fulfill

the Requirements for the Degree of
„doctor rerum naturalium“ (Dr. rer. nat.)

**Submitted to the Council of the Faculty of
Biological Sciences
of the Friedrich Schiller University Jena**

by **Sagar Changdev Gore** (Master of Science)

born on **26th July 1991** in **Khatav, India**

Reviewers:

Prof. Dr. Dirk Hoffmeister

Pharmaceutical Microbiology

Friedrich Schiller University, Jena.

Prof. Dr. Stefan Schuster

Department of Bioinformatics

Friedrich Schiller University, Jena.

Jun-Prof. Dr. Panagiotis L. Kastiris

Group Leader and Head of Cryo-Electron Microscopy

Martin-Luther-Universität Halle-Wittenberg, Biozentrum, Halle (Salle).

Date of doctoral defense: June 16th, 2020.

Summary	8
Zusammenfassung	10
Table of contents	
1. Introduction	15
1.1 Why is there a need for new drugs?	16
1.1.1 Antibiotic resistance	16
1.2 Natural products-based drug discovery	16
1.3 Fungi – a source of novel natural products	18
1.4 Why are fungal natural products not studied well?	20
1.5 Non-ribosomal peptide synthetases (NRPS)	21
1.6 Adenylation domains	22
1.7 Substrate specificity of A-domains	24
1.8 Condensation domains	27
1.9 Encoding of small molecules	28
1.9.1 Molecular representation	28
1.9.2 Molecular descriptors	29
1.9.3 Molecular fingerprints	29
1.9.4 Molecular similarity	30
1.10 Computational methods for A-domain substrate specificity predictions	31
1.10.1 SANDPUMA: an ensemble classifier	31
1.10.2 Limitations and drawbacks of the previously developed tools	33
1.11 Machine learning	33
1.11.1 Artificial neural networks	34
1.11.2 Evaluation of predictor performance	36

2. Methods	40
2.1 A-domain substrate specificity data	40
2.2 Phylogenetic analysis	40
2.2.1 Phylogenetic analysis of fungal A-domain sequences	40
2.2.2 Extraction of C-domains and phylogenetic analysis	41
2.3 Clustering of bacterial and fungal A-domain NRPS codes	41
2.4 Clustering of non-ribosomal peptide monomers	41
2.5 Development of NN-based A-domain substrate specificity classifier (NNasc)	42
2.5.1 Dataset preparation for NNasc	42
2.5.2 Preprocessing	43
2.5.2.1 Encoding of NRPS code residues	43
2.5.2.2 Encoding of substrates	44
2.5.3 Parameter optimization and training NNasc	44
2.5.4 Matching of predicted bit vector with substrates from training dataset	45
2.5.5 Validation and benchmarking	46
2.5.5.1 Internal validation dataset	46
2.5.5.2 External validation dataset	46
3. Results and discussion	49
3.1 Phylogenetic analysis of fungal A-domains	49
3.1.1 A-domains activating uncommon non-proteogenic substrates	49
3.1.2 ACVS synthetase	49
3.1.3 PKS-NRPS hybrids	49
3.2 Cluster analysis of NRPS code residues	53
3.2.1 Similarity/differences between fungal and bacterial NRPS codes	53
3.3 Phylogenetic relationships among fungal C-domains	56

3.4 Clustering of NRP monomers and A-domain substrates	57
3.4.1 Modifications	57
3.4.2 A-domain substrates and NRPS code similarity	59
3.5 Substrate specificity prediction using NNasse	60
3.5.1 Predictions – internal validation dataset	61
3.5.2 Predictions - external validation dataset	61
3.5.3 Structural insights into substrate specificity prediction	63
3.5.4 Prediction of substrate bit vectors	63
3.5.5 Inclusion of more training data	64
3.5.6 Extraction of correct NRPS code residues	64
4. Conclusions	66
5. References	68
Appendix	75
Acknowledgement	82
Curriculum Vitae	84
List of publications	86
Declaration/ Selbstständigkeitserklärung	88

Summary

Non-Ribosomal Peptide Synthetases (NRPS) are mega synthetases that are predominantly found in bacteria and fungi. They produce small peptides that serve numerous biological functions and crucial ecological roles. NRPSs are organized into different modules, each responsible for processing (substrate activation or chemical modifications) and incorporation of single substrate monomers. A single module of NRPS has three domains such as Adenylation (A), Condensation (C), and Thiolation (T). Substrate selection and their activation by adenylation is a key step in the non-ribosomal peptide (NRPs) biosynthesis. A-domains catalyze ATP dependent activation of substrates harboring carboxy terminus. A-domain substrates include not only natural amino acids (D and L forms) but also non-proteinogenic amino acids. The inclusion of non-proteinogenic amino acid increases the structural and chemical diversity of biosynthesized peptides. C-domains have also been reported to show moderate specificity towards a substrate that is activated by an A-domain. Norine database has listed 543 monomers (247 NRPS specific) in their repository, these monomers are extracted from the complete chemical structures of characterized NRPs. As the substrate repertoire is large and specificity rules for fungi are not established well, there is a difficulty in predicting substrates for fungal A-domains. In bacteria, ten amino acid residues were established as “NRPS code”, which determine specificity of A-domains. These residues were identified by mutagenesis studies done with a phenylalanine binding A-domain.

To study relationships between fungal A-domains and their specificity, I ran the cluster analysis of NRPS code residues. Fungal A-domain sequences with known substrates and corresponding NRPS code residues were obtained from the NRPSpredictor2 dataset (released in 2011). This dataset was expanded by adding manually curated sequences that were published after 2011. NRPSpredictor2 is A-domain substrate specificity classifier, which uses support vector machine models for the prediction. NRPS code residues were encoded by physicochemical properties essential for binding small molecules and these residues were clustered by their similarity. Cluster analysis showed similar NRPS codes for α -amino adipic acid, phenylalanine, and tryptophan, etc. between bacteria and fungi. Fungal NRPS codes for substrates such as tyrosine, serine, and proline, did not cluster together with bacteria, which indicates an independent evolution of substrate specificity in fungi. This emphasizes the pressing need for the development of a fungus-specific prediction tool. Currently available A-domain sequence-based specificity prediction tools accurately identify

substrates for bacteria but fail to provide correct predictions for fungi. SANDPUMA, an ensemble classifier for A-domain substrate specificity prediction, developed in 2017, uses only 90 fungal sequences (complete dataset – 928 sequences) but does not give accurate substrate predictions for fungal sequences.

I present here a novel approach for fungal A-domain substrate specificity prediction, which is based on a neural network (NN). I developed the NN-based A-domain substrate specificity classifier (NNassc) using Keras with TensorFlow backend using Python scripting language. It was trained solely using fungal NRPS codes and combines physicochemical and structural features for specificity predictions. Internal and external validation datasets of experimentally verified NRPS codes were used to assess the performance of NNassc. NNassc predictions were compared with SANDPUMA using these two validation datasets. As opposed to earlier sequence-based prediction tools, our approach involves the prediction of substrate substructures rather than mere substrate classes. NNassc identifies correct substrates within the top three or five predictions in all cases except for phenylalanine, while SANDPUMA works only with alanine or tyrosine NRPS codes. The substrate is encoded as a Morgan fingerprint bit vector, such that each bit encodes certain molecular property or a substructure. Prediction of substructures has an advantage that the novel (not part of the training dataset) substrates could also be predicted that harbor these substructures. In the case of NNassc, comparisons of predicted fingerprint bit-vectors were done only with training dataset substrates, although a larger database (e.g. Norine database) of probable substrates could be used in difficult cases. In the future, A-domain substrate specificity predictions could be improved by including more NRPS codes for each substrate and structural information about specific binding interactions.

Zusammenfassung

Nicht-Ribosomale Peptidsynthesen (NRPS) sind Megasyntetasen, die überwiegend in Bakterien und Pilzen vorkommen. Sie produzieren kleine Peptide, die zahlreiche biologische Funktionen und wichtige ökologische Funktionen erfüllen. NRPSs bestehen aus verschiedenen Modulen, die jeweils für die Verarbeitung (Substrataktivierung oder chemische Modifikationen) und die Einbindung einzelner Substratmonomere verantwortlich sind. Ein einzelnes NRPS-Modul besteht aus drei Domänen: Adenylierung (A), Kondensation (C) und Thiolyierung (T). Die Substratselektion und seine Aktivierung durch Adenylierung ist ein wichtiger Schritt in der Biosynthese von nicht-ribosomalen Peptiden (NRPs). A-Domänen katalysieren die ATP-abhängige Aktivierung von Substraten mit Carboxyterminus. A-Domänen-Substrate umfassen nicht nur natürliche Aminosäuren (D- und L-Form), sondern auch nicht-proteinogene Aminosäuren. Das Zulassen von nicht-proteinogenen Aminosäuren erhöht die strukturelle und chemische Vielfalt der biosynthetisierten Peptide. Es wurde auch berichtet, dass C-Domänen eine moderate Spezifität gegenüber einem Substrat aufweisen, das durch eine A-Domäne aktiviert wird. Die Norine-Datenbank hat 543 Monomere (247 NRPS spezifisch) in ihrem Repositorium aufgelistet. Diese Monomere wurden aus den kompletten chemischen Strukturen charakterisierter NRPs extrahiert. Da das Substratrepertoire groß ist und Spezifitätsregeln für Pilze nicht ausreichend bekannt sind, gibt es Schwierigkeiten bei der Vorhersage von Substraten für Pilz A-Domänen. In Bakterien wurden zehn Aminosäurereste als "NRPS-Code" etabliert, die die Spezifität für A-Domänen bestimmen. Diese Aminosäurereste wurden durch Mutagenese-Studien einer Phenylalanin-Bindungs A-domäne identifiziert.

Um den Zusammenhang zwischen Pilz-A-Domänen und ihrer Spezifität zu untersuchen, führte ich eine Clusteranalyse von NRPS-Code-Resten durch. Pilz-A-Domänen-Sequenzen mit bekannten Substraten und entsprechenden NRPS-Code-Resten wurden aus dem NRPSpredictor2-Datensatz (entwickelt 2011) bezogen. Ich habe nach 2011 entdeckte Sequenzen manuell kuratiert und dem NRPSpredictor2-Datensatz hinzugefügt. NRPSpredictor2 ist ein Substratspezifitätsklassifizierer für die A-Domäne, der zur Vorhersage Support-Vektor-Maschinenmodelle verwendet. NRPS-Code-Reste wurden durch physikalisch-chemische Eigenschaften codiert, die für die Bindung kleiner Moleküle wesentlich sind, und diese Reste wurden aufgrund ihrer Ähnlichkeit geclustert. Die Clusteranalyse zeigte ähnliche NRPs-Codes für α -Aminoadipinsäure, Phenylalanin und Tryptophan, etc. zwischen Bakterien und Pilzen. Pilz NRPS-Codes für Substrate wie Tyrosin, Serin und Prolin

bildeten keine Cluster mit Bakterien, was auf eine unabhängige Evolution der Substratspezifität in Pilzen hindeutet. Dies unterstreicht die dringende Notwendigkeit der Entwicklung eines pilzspezifischen Vorhersagetools.

Derzeit verfügbare sequenzbasierte A-Domänen-Spezifitätsvorhersageprogramme identifizieren Substrate für Bakterien genau, liefern aber keine korrekten Vorhersagen für Pilze. SANDPUMA, ein 2017 entwickelter Ensemble-Klassifizierer für die Vorhersage der Substratspezifität in der A-Domäne verwendet nur 90 Pilzsequenzen (vollständiger Datensatz - 928 Sequenzen) und liefert keine genauen Substratvorhersagen für Pilzsequenzen.

Ich präsentiere hier einen neuartigen Ansatz zur Vorhersage der Substratspezifität von Pilz-A-Domänen, der auf einem neuronalen Netzwerk (NN) basiert. Ich habe einen NN-basierten A-Domain-Substratspezifitätsklassifizierer (NNasc) unter Verwendung von Keras mit TensorFlow-Backend in der Python-Skriptsprache entwickelt. Er wurde ausschließlich unter Verwendung von NRPS-Codes für Pilze trainiert und kombiniert physikalisch-chemische und strukturelle Merkmale zur Vorhersage der Substratspezifität. Interne und externe Validierungsdatensätze von experimentell verifizierten NRPS-Codes wurden verwendet, um die Performanz von NNasc zu beurteilen. NNasc-Vorhersagen wurden mit SANDPUMA anhand dieser beiden Validierungsdatensätze verglichen. Im Gegensatz zu früheren sequenzbasierten Vorhersagewerkzeugen beinhaltet unser Ansatz die Vorhersage von Substratunterstrukturen und nicht nur von Substratklassen. NNasc identifiziert in allen Fällen mit Ausnahme von Phenylalanin korrekte Substrate innerhalb der ersten drei oder fünf Vorhersagen, während SANDPUMA nur mit Alanin- oder Tyrosin-NRPS-Codes arbeitet. Das Substrat wird als Morgan-Fingerabdruck-Bitvektor kodiert, so dass jedes Bit eine bestimmte molekulare Eigenschaft oder eine Teilstruktur kodiert. Die Vorhersage von Teilstrukturen hat den Vorteil, dass auch neuartige (nicht Teil des Trainingsdatensatzes) Substrate vorhergesagt werden konnten, die diese Teilstrukturen besitzen. Im Falle von NNasc wurden Vergleiche der vorhergesagten Fingerabdruck-Bitvektoren nur mit Trainingsdatensatz-Substraten durchgeführt, obwohl eine größere Datenbank (z. B. Norine-Datenbank) mit wahrscheinlichen Substraten in schwierigen Fällen verwendet werden könnte. In Zukunft könnten die Vorhersagen der Substratspezifität in der A-Domäne verbessert werden, indem mehr NRPS-Codes für jedes Substrat und strukturelle Informationen über spezifische Bindungsinteraktionen einbezogen werden.

“Trying to understand the way nature works involves a most terrible test of human reasoning ability. It involves subtle trickery, beautiful tightropes of logic on which one has to walk in order not to make a mistake in predicting what will happen. The quantum mechanical and the relativity ideas are examples of this”.

Richard P. Feynman

(“The Uncertainty of Science”, John Danz Lecture Series, 1963)

1. Introduction

Secondary metabolites or natural products are specialized metabolites produced by different (micro) organisms, which gives them survival advantages in their respective environmental niches. These metabolites play multifarious roles (biological, ecological, pharmaceutical and agricultural purposes) and hence these could be exploited for human use. Nature has endowed us with diverse sources of these metabolites, namely bacteria, fungi, plants and marine animals (metabolites were produced by microbial symbionts, Unson *et al.*, 1993). Recent metagenomic studies have helped unearth many unimagined and previously unexplored sources of natural products (NPs), e.g., the human gut microbiome (Donia *et al.*, 2014), marine samples, etc. (Hyde *et al.*, 2019). Besides, NPs crucial ecological role, they have been important for the humankind in pharmacological (the inspiration for drugs) and agricultural (insecticides, fungicides) areas. “The golden age of antibiotic discovery” was termed for the period of the 1930s till the 1950s, during which the potent antibiotics were discovered. It all started with the serendipitous discovery of penicillin by Alexander Fleming in 1928. Traditional drug discovery after this era involved time-consuming steps of isolation and purification of small molecules that modulate biological targets. Target-based (specific biological target) screening was practiced for a long time, but with the advent of gene sequencing technologies and usage of computational chemistry, this changed. During the 1980s to early 1990, many pharmaceutical companies moved from NP based drugs to synthetic compounds. It was also the time when molecular biology techniques were gaining widespread attention. In recent years NPs have received renewed attention from many academic research labs for novel compounds for drug discovery, although finding chemically distinct scaffolds is still a big challenge. Despite technological advances and a better understanding of disease mechanisms, there are still low success rates in drug discovery programs.

For this thesis, the main goal was to predict substrates for fungal adenylation domains, which are key enzymes of non-ribosomal peptide synthetase (NRPS) involved in substrate selection. NRPSs are mega synthetases for secondary metabolite biosynthesis, which are generally organized into different modules. The neural network-based machine learning model for the prediction of substrates for adenylation domains was developed and validated. With this model, it would be possible to assign substrates to fungal A-domains from sequenced genomes. Consequently, it helps to move one step closer to predict the complete structure of fungal non-ribosomal peptides.

In the following sections, I shall give a brief overview of NPs and their potential in drug discovery. I shall also describe molecular representation for small molecules and introduce machine learning methods in the upcoming sections, which will be useful later for the development of the model.

1.1 Why is there a need for new drugs?

Many neglected tropical diseases e.g. African trypanosomiasis (sleeping sickness), leishmaniasis, lymphatic filariasis, schistosomiasis have limited treatment options, hence more research for the development of novel treatments is warranted. One of the other crucial reasons for discovering or designing novel drugs is to combat infections caused by microbes, as they are becoming highly resistant to currently available drugs.

1.1.1 Antibiotic resistance

Antibiotic resistance is the ability by which bacteria acquire or develop resistance against previously effective drugs. Horizontal transfer of antibiotic resistance genes is one of many mechanisms by which microbes become resistant (Von Wintersdorff *et al.* 2016) (Table 1). These resistance genes are acquired from the bacteria that produce these antibiotics (Martinez, J. L. 2014, Reygaert W. 2018). The second reason for the prevalence of antibiotic resistance is their widespread and haphazard use. To circumvent this problem of resistance, there is an urgent need for novel drugs or therapies.

1.2 Natural products-based drug discovery

Drug discovery must fulfill two prime conditions that, the drugs are safe and effective. Accomplishing these conditions relies on a range of other factors namely bioavailability, absorption, distribution, metabolism, excretion, and toxicology, which are also termed as ADMET properties. Drug discovery involves laborious steps of from the target identification up to the clinical trials; hence it is a long and expensive endeavor. Computational efforts help accelerate this process and effectively reduce the monetary burden. Recent efforts towards computational drug discovery encompass structure-based, ligand-based, or fragment-based drug design approaches (Macalino *et al.*, 2015). Structure-based drug design methods have been synergistically used with other experimental methods for the development of few drugs (e.g. Boceprevir is used as a protease

inhibitor to treat hepatitis). Some of these drugs are at different stages of clinical trials (Talele *et al.*, 2010).

Table 1. List of antibiotics and year of their development and year when resistance was first observed in bacteria. Source: Centers for disease Control and Prevention.

Antibiotic (year introduced)	Year resistance was reported (bacteria)
Penicillin (1943)	1965 (Penicillin-R Pneumococcus)
Tetracycline (1950)	1959 (Tetracycline-R Shigella)
Erythromycin (1953)	1968 (Erythromycin-R Streptococcus)
Methicillin (1960)	1962 (Methicillin-R Staphylococcus)
Gentamicin (1967)	1979 (Gentamicin-R Enterococcus)
Vancomycin (1972)	1988 (Vancomycin-R Enterococcus), 2002 (Vancomycin-R Staphylococcus)
Imipenem and Ceftazidime (1985)	1987 (Ceftazidime-R Enterobacteriaceae)
Levofloxacin (1996)	1996 (Levofloxacin-R Pneumococcus)
Linezolid (2000)	2001 (Linezolid-R Staphylococcus)
Ceftaroline (2011)	2011 (Ceftaroline-R Staphylococcus)

NPs possess certain physicochemical properties that make them promising candidates for drug discovery. These properties vary depending upon the class of compounds. Some classes of the studied NPs show many chiral centers, *sp*³ configured atoms, characteristic 3D shape, higher molecular complexity, lower hydrophobicity, larger molecular scaffolds than synthetic drugs which are crucial properties to act as potential drug candidates (Stratton *et al.*, 2016). Many NPs follow the rule of five (Lipinski *et al.*, 1997), hence they are good starter molecules for drug design. Many US

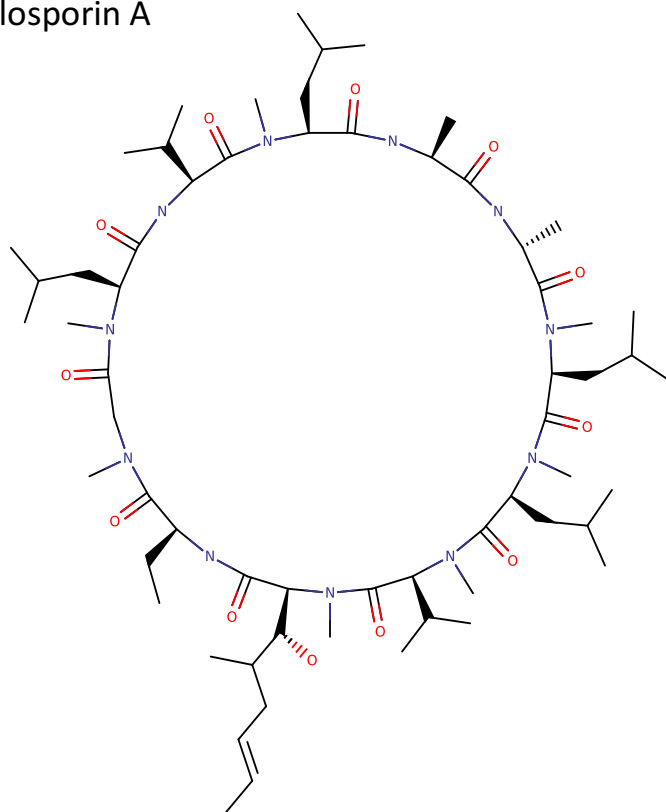
food and drug administration-approved drugs are either derived from NPs or NPs themselves (Patridge *et al.*, 2016), e.g. in case of drugs approved for cancer (between the 1940s - 2014) – 49% are NPs or direct derivatives (Newman *et al.*, 2016). NPs harbor substructures or functional groups that are optimized through the course of evolution to selectively bind biological targets. Computational tools such as ChemGPS-NP (Larsson *et al.* 2007), ScaffoldHunter (Wetzel *et al.* 2009) have been pivotal in finding NP derived fragments (Rekar *et al.* 2014). NP derived fragments have been successfully used as inhibitors for p38a MAP kinase of by Waldmann and coworkers (Björn *et al.* 2013). NPs could be used to selectively target, proteins of interest, as they are less promiscuous in comparison to synthetic drugs (Schneider *et al.*, 2016, Tiago *et al.*, 2016). Peter Ertl and coworkers compared NPs produced by different organisms and they found out that, NPs produced by fungi and plants are most closely related. This is because the NPs produced by fungi and plants share more functional groups with each other than with bacteria (Ertl *et al.*, 2019).

1.3 Fungi – a source of novel natural products

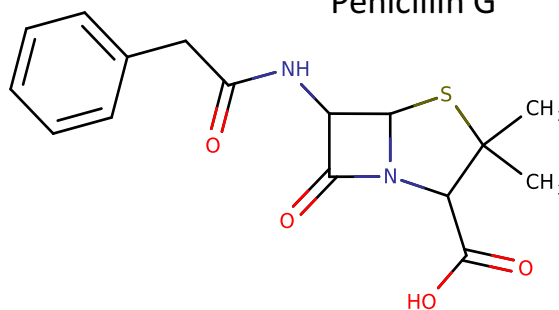
Fungi produce an array of structurally and chemically diverse NPs that could have detrimental or beneficial effects for humans or other organisms. These different classes include but are not limited to polyketides, ribosomal and non-ribosomal peptides, terpenes, alkaloids, etc. The structures of few fungal non-ribosomal peptides are shown in Fig1. Precursors for NP biosynthesis are obtained either from primary metabolism or from pathways dedicated to their synthesis. Genes involved in NP synthesis are often arranged in one contiguous genetic locus, which is called the biosynthetic gene cluster (BGC). BGCs are comprised of a set of genes that are co-localized (on chromosome) and are co-expressed under certain conditions. BGCs are composed of genes for core enzymes responsible for the complete biosynthesis of NPs from monomers, as well as genes for accessory enzymes. In addition to this, genes for transcription factors (for the regulation of expression), transporters (for detoxification of compounds) could also be a part of BGCs (Keller *et al.*, 2019). Minimum Information about a Biosynthetic Gene Clusters (MIBiG), a systematic and well-curated resource and a standard for deposition of BGCs (in bacteria, fungi, and algae) was conceptualized by Medema *et al.*, 2015. Multiple studies have suggested that fungal genomes are replete with cryptic BGCs without any knowledge of encoded compounds. Li *et al.*, 2016 collected experimentally characterized 197 NPs, and they highlight that only certain compound classes e.g. Aflatoxin or fungal phyla e.g. *Aspergilli* are overrepresented in these. Specifically, fungal species of Basidiomycota are

less well characterized, and these studies point towards their tremendous untapped potential to produce novel NPs (Lackner *et al.*, 2012, Li *et al.*, 2016, Stadler *et al.*, 2015, Frisvad *et al.*, 2017).

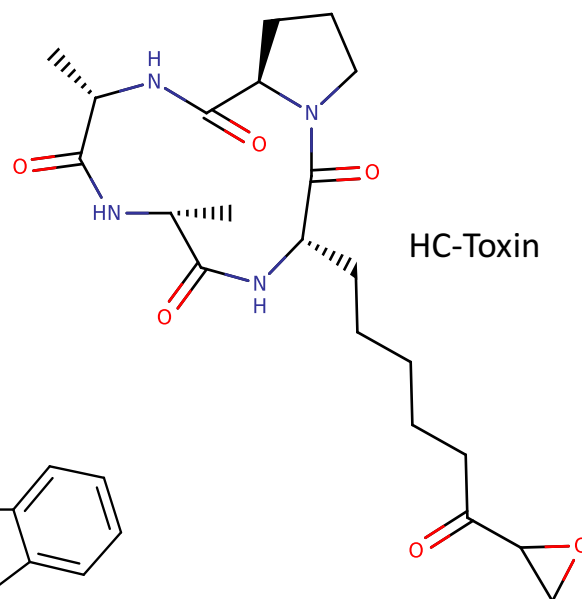
Cyclosporin A



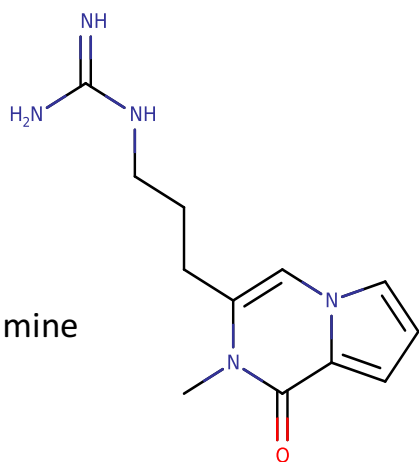
Penicillin G



HC-Toxin



Peramine



Tryptoquialanine

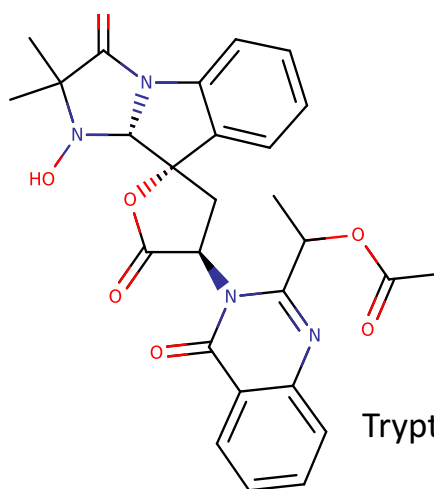


Fig 1. Chemical structures for fungal NPs. Cyclosporin A, Penicillin G, Peramine, HC-toxin, and Tryptoquialanine. Structures were edited in MarvinSketch (version 19.24, developed by ChemAxon, <http://www.chemaxon.com>) and SDF files were obtained from the PubChem database.

1.4 Why are fungal natural products not studied well?

The numbers of bacterial genomes sequenced is manifold higher than in fungi, hence comparative analysis of bacterial sequences to study their biosynthetic potential is possible. While many bacterial BGCs are well studied, characterization of fungal BGCs and their products lags well behind their actual capacity. The difficulty to study fungal biosynthetic machinery arises due to multifarious factors such as slower growth rates, lower product yields, inactive BGCs, etc. Some of the strategies to study fungal NP repertoire are changing growth conditions, providing osmotic or oxidative stress, the addition of elicitors, co-culture with other microorganisms, and heterologous expression of BGC genes (Hoefgen *et al.*, 2018). However, many fungi are not amenable to genetic manipulations. All these factors make NP characterization in fungi much slower and less intensive than that in bacteria.

Also, the experimental approaches are time and labor-intensive, and they do not guarantee success in discovering novel compounds. Many a time, same NPs are rediscovered after experimental characterization of seemingly novel BGCs. This calls for an alternative, yet powerful approach of using state-of-the-art computational methods to mine fungal genomes for new chemical structures. In last few years, computational prediction of NPs followed by systematic experimental characterization has become commonplace because of the development of powerful bioinformatics algorithms e.g. antiSMASH 4.0 (identification of BGCs in bacteria and fungi and prediction of encoded NPs) (Blin *et al.*, 2017), CASSIS (BGC prediction in genomes of eukaryotes) (Wolf *et al.*, 2016), SMURF (prediction of BGCs in fungi) (Khaldi *et al.*, 2010), PhytoClust (identification of BGCs in plants) (Töpfer *et al.*, 2017), PRISM 3.0 (prediction of chemical structures of NRPs) (Skinnider *et al.*, 2017). The development of better tools for fungal and plant genome sequence analysis are needed, as previously developed tools use bacterial data to train their algorithms.

In the following sections, Non-ribosomal peptide synthetases and roles of individual domains (mainly adenylation and condensation) in the biosynthesis of peptides are discussed. This lays the groundwork for the main research work of this thesis i.e. substrate specificity prediction of adenylation domains.

1.5 Non-ribosomal peptide synthetases (NRPS)

NRPSs are megasynthetases that biosynthesize non-ribosomal peptides (NRPs) from small building blocks. NRPSs are composed of one or more self-sufficient modules required for selective activation of amino acids and their incorporation into the peptide chain. The minimal module consists of an adenylation (A), thiolation (T) or peptidyl carrier protein (PCP) and condensation (C) domains. A-domains are capable of specifically selecting substrates and activating them by adenylation. These domains transfer activated amino acids to a phosphopantetheine moiety attached to T-domain. This moiety is covalently linked to conserved serine hydroxy groups in T-domain. C-domains are involved in condensation of substrates activated by A-domains. Although the substrate selection is primarily done by A-domains, C-domains are known to show moderate specificity towards substrates activated by downstream A-domains (Rausch *et al.*, 2007). NRPSs also contain accessory domains, which are involved in modifying the growing peptide or monomers or substrates. For e.g. Epimerization (E) domain inverts the stereochemistry of the alpha carbon of the substrates amino acids, methyltransferase (M) domain transfers methyl group on oxygen or nitrogen. The terminal domain of NRPS is responsible for the catalytic release of the complete peptide. Thioesterase (TE) and cyclization (Cy) or condensation like (CT) domains are usually found in bacterial and fungal NRPSs respectively (Fig 2.).

The modular nature of the NRPS mega synthetases makes them promising candidates for the engineering of desired compounds. There are multiple ways of the engineering of NRPSs such as directed evolution, random mutagenesis, single residue mutations or module recombination. Although each of them is associated with their own set of challenges, to overcome these, further understanding of the NRPSs biosynthesis mechanism is required. One of the important steps is to better understand the fine mechanisms of A- and C-domain substrate specificities.

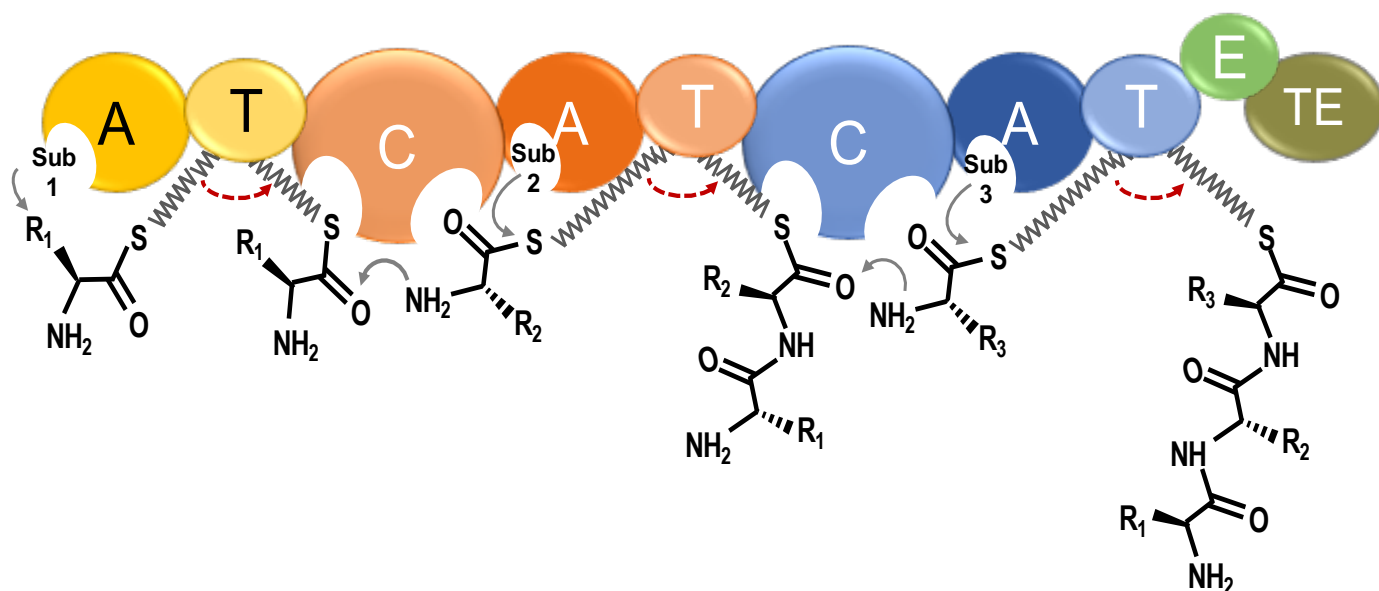


Figure 2. Schematic representation for NRPS mega synthetase biosynthetic mechanism. A-domains select substrates and activate them by adenylation reaction, C-domains possess two binding sites for substrates activated by adjacent A-domains. T-domain tethers activated substrate and the growing peptide. A-Adenylation, T-Thiolation, C- Condensation, and TE- Thioesterase domain. NRPSs are terminated usually by TE and CT domains in bacteria and fungi respectively except in ACVS synthetases found in fungi have TE domains at their termini.

1.6 Adenylation domains

NRPS A-domains are a part of ANL (Acyl-CoA synthetase, Non-ribosomal peptide synthetase, Luciferase) family which belongs to adenylate forming enzyme superfamily. Other members of this superfamily include class I and II aminoacyl-tRNA synthetases, which are involved in ribosomal peptide synthesis. They have different structural folds compared to NRPS A-domains but catalyze the same biochemical reaction.

A-domains have large N-terminal and smaller C-terminal subdomains. A-domains catalyze two enzymatic reactions, the first one being recognition and activation of the carboxyl group-containing substrates by adenylation reaction and the second, transfer of acyl-AMP intermediate to the

phosphopantetheine arm of the T (or PCP) domain. A-domains are crucial as they catalyze ATP dependent activation of otherwise unreactive carboxylic acids. Mechanism of substrate recognition and activation by adenylation reaction has been studied by structural analysis of individual (or in complexes) domains for NRPSs. A-domain 3D structures that are deposited in Research Collaboratory for Structural Bioinformatics Protein data bank (RCSB PDB) are listed in Table 2.

Table 2. NRPS A-domain three dimensional structures deposited in RCSB PDB database. PDB ID and bound substrate information is given along with year of structure determination.

PDB ID	Protein	Secondary metabolite	Bound substrate	Year
3ITE	SidNA3	Siderophore	N(δ)-cis-anhydromevalonyl-N(δ)-hydroxy-L-ornithine	2010
6P4U, 6P3I, 6OYF, 6OZV, 6P1J	Txo1	Teixobactin	-	2019
5N82	TycA	-	(S)-beta-phe	2018
5WMM	ThioS	Thiocoraline	Norcoronamic acid (and L-val)	2018
5N9W, 5N9X	Thr1	Chloro Thr	L-Thr	2017
5JJQ	IdnL1	Incednine	3S-3-aminobutyic acid	2017
5JJP	CmiS6	Cremimycin	3-aminononanoic acid	2017
5ES8	LgrA	Linear gramicidin	L-Val	2016
4ZXI	AB3404	Unknown	Gly	2016
5T3D	EntF	Enterobactin	L-Ser	2016
4WV3	AuaEII	Aurachin	Anthranilic acid	2016

4R0M	McyG	Microcystin	L-Phe	2015
4D56, 4D57	ApnA-A1	Anabenopeptin	L-Arg/L-Tyr	2015
4OXI	AlmE	LPS modification	Gly	2014
3WV5	VinN	Vicenistatin	Beta-methyl-L-Asp	2014
4GR5	SlgN1	Streptolydigin	Beta-methyl-L-Asp	2013
4DG8, 4DG9	PA1211	Unknown	L-Val	2012
2VSQ	SrfA-C	Surfactin	L-Leu	2008
3DHV	DLTA	D-alanylation of lipoteichoic acid	D-Ala	2008
3VNR, 3VNQ, 3VNS	CytC1	Cytotrienin	2-aminobutyric acid	2007
1AMU	PheA	Gramicidin	L-Phe	1997

1.7 Substrate specificity of A-domains

A-domains are known to selectively activate substrates by coordinating with them through a set of binding site residues. Marahiel and coworkers (Stachelhaus *et al.*, 1999) defined the 10 residues involved in recognition of substrates as “NRPS code” (Fig 3). These residues were deduced from bacterial phenylalanine activating A-domain (PheA). Mutations of NRPS code residues in PheA are shown to alter the substrate specificity to activate non-cognate substrates. Pyrophosphate (PPi) exchange assay with a highly purified enzyme is a standard method to elucidate substrates activated by A-domains.

A-domains recognize not only twenty natural amino acids but other non-proteogenic acids such as fatty acids, hydroxy acids, aromatic acids, aryl acids, keto acids, etc. Recent studies by De Mattos-

Shiple et al. 2018 show that methylated amino acids could also be accepted by A-domains. Incorporation of these non-proteinogenic amino acids into NRP provides increased structural diversity, protease degradation resistance, and stereochemical constraints. A-domain substrates and NRP monomers that are incorporated in NRPs are deposited in the Norine database by Pupin and coworkers (Flissi *et al.*, 2015). Although this database does not include all A-domain substrates characterized, 247 A-domain monomers pertinent to NRPS mega synthetases are listed there. Accessory enzymes modify monomers or growing peptides, hence monomers deduced from the final peptide structure could be different from substrates activated by A-domains.

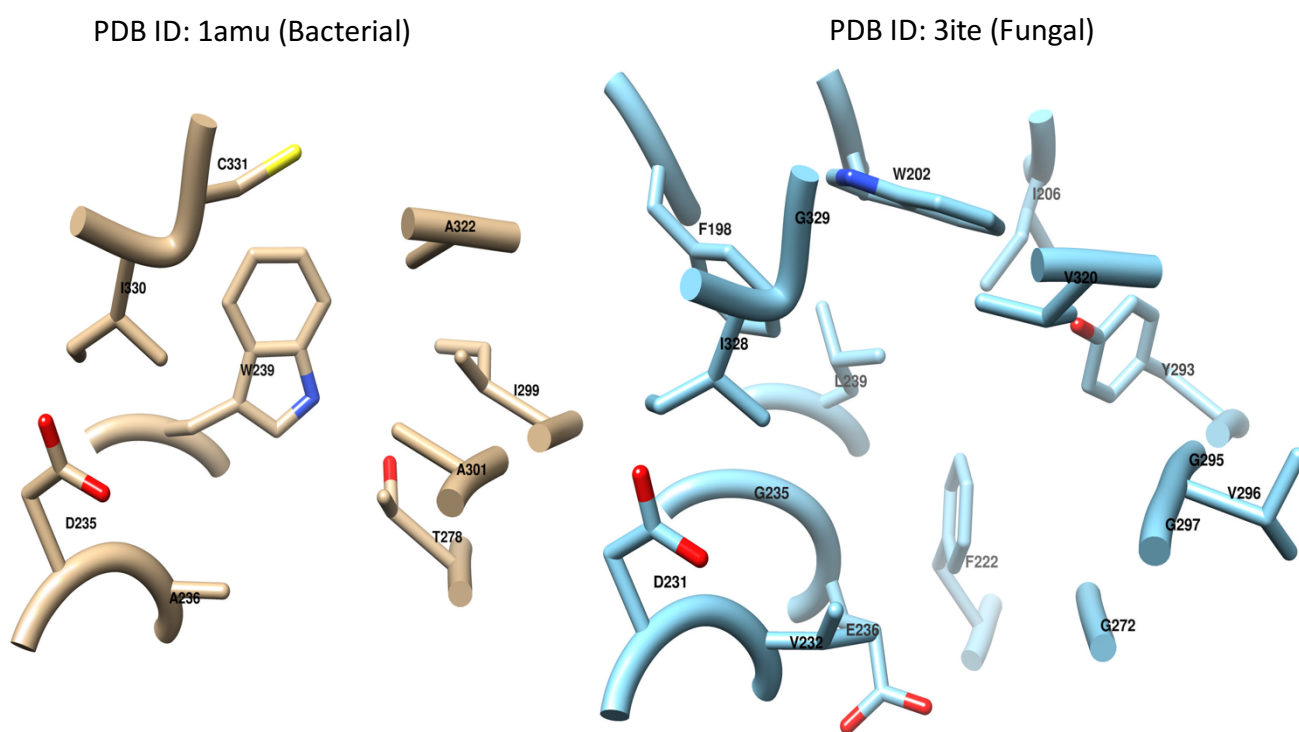


Figure 3. NRPS code residues from bacterial and fungal A-domains. Nine residues for the bacterial PheA (binds phenylalanine) NRPS code is shown (highly conserved catalytic lysine residue is not shown). 17 residues from fungal SidNA3 (binds large substrate, N(δ)-cis-anhydromevalonyl-N(δ)-hydroxy-L-ornithine) are shown. W239 from PheA at the base of pocket controls the pocket size (in fungi this residue is replaced by G235).

Substrate specificity differences in A-domains could be brought about by mechanisms such as sequence divergence, duplication, recombination events or sequence polymorphisms. A-domains

share sequence identity in the twilight zone (20-35%) with its homologs, hence there is a difficulty in building reliable 3D structural models. Despite high sequence divergence on the overall low sequence identity background, 3D structural folds are well conserved. Lee Verena and coworkers employed A-domain 3D modeling and ligand-docking based approach for substrate specificity inference, but they emphasized that the accuracy of models built was a major bottleneck to predict substrates correctly (Lee *et al.*, 2015).

A-domains are the gatekeepers for substrates incorporated into growing NRPs. The logic behind the selection of specific substrates lies in substrate binding residues. Structural elucidation of A-domains and other NRPS proteins through x-ray crystallography, nuclear magnetic resonance or cryo-EM microscopy is possible but is not always straightforward. In such difficult cases, computational approaches could be used to predict A-domain substrates and relevant protein-protein interactions.

Some residues of NRPS code coordinate directly with the substrate, while others are involved in positioning directly interacting residues for optimal interactions (Lee *et al.* 2010). Substrates could have directional hydrogen bond interactions with polar residues or hydrophobic interactions with aliphatic or aromatic residues. There are multiple ways of binding the same substrate, which is exemplified by the presence of many non-redundant NRPS codes for the same substrates. These differential interactions cannot be identified without experimentally solved structures of A-domain-substrate complexes. Out of all A-domain 3D structures deposited in the RCSB PDB database, only one is of eukaryotic origin (PDB ID: 3ITE): the third A-domain of the fungal SidN with specificity to N(δ)-cis-anhydromevalonyl-N(δ)-hydroxy-L-ornithine (Lee *et al.* 2010). In contrary to ten NRPS code residues extracted from bacterial A-domain (substrate - phenylalanine), fungal SidN suggests 17 binding residues to interact with an unusually large substrate (L-ornithine derivative).

Depending on the size and locations of charged residues of a binding pocket, different substrates can be bound and activated by the same pocket. For example, tyrosine and arginine which are structurally, and chemically different substrates could bind to the same ApnA A1 pocket (Kaljunen *et al.* 2015). Crystal structures (PDB IDs: 4D56, 4D57) of bound substrates depict identical binding shapes and interactions for both substrates. These insights are not possible without high-resolution structures of A-domain-substrate complexes. A structural basis for non-proteogenic amino acid binding NRPS codes was recently studied (Kudo *et al.*, 2018). These studies reiterate the importance of structural information to deduce A-domain substrate recognition rules.

1.8 Condensation domains

Condensation (C) domains catalyze amide bond formation between two amino acids (or hydroxy or aryl acids) in a growing peptide chain. C-domains have two binding sites, one for aminoacyl acceptor and the other for binding a peptidyl donor. C domains are suspected to favor only some substrates activated by A-domains and hence imparting one more layer of specificity. Rauch *et al.*, 2007 categorized bacterial C-domains based on stereochemistry of substrates condensed (L or D- amino acids) or accessory chemical reactions catalyzed. These groups are LCL, DCL, Starter C domains, heterocyclization (Cyc) and, dual Epimerization/Condensation domains. These C-domains or C-domain like enzymes catalyze other chemical reactions in place or along with condensation reaction (Kraas *et al.*, 2012; Linne and Marahiel, 2000; Rausch *et al.*, 2007; Stachelhaus and Marahiel, 1995). C-domains harbor catalytic motif HHXXXDG, residue mutations from this motif have shown to alter condensation activity (Bergendahl *et al.*, 2002). Bloudoff *et al.*, 2016 provided the first C-domain acceptor substrate analog bound structure (PDB ID: 5DU9 and 5DUA), which is obtained by covalent tethering this analogue near the active site. The acceptor substrate α -amino group acts as an H bond donor and interacts with H157 and S386. Mutational analysis performed in this work also suggested that H157 plays a key role in substrate positioning and S309 imparts acceptor substrate selectivity.

In bacteria, thioesterase (TE) domains are dedicated to cyclization of peptide precursors, while in fungi, C-domain like CT domains catalyze the release of peptide products or their macrocyclization. Phylogenetic analysis suggests that these CT domains show higher similarity among different multi and mono modular NRPSs than with C-domains from the same NRPS (Haynes *et al.*, 2014). Cyclosporine, tryptoquialanine or fumiquinazoline (alkaloids) are some compounds whose NRPSs harbor CT domains at the end.

There have been several attempts to engineer NRPS assembly lines, but there were difficulties to produce desired compounds in higher amounts (Zobel *et al.*, 2016, Winn *et al.*, 2016). This could be due to an incomplete understanding of and A and C-domain substrate specificities (Rausch *et al.*, 2007) and protein-protein interactions among various domains. Helge Bode and coworkers introduced a novel concept of exchange units (XUs) to engineer NRPS assembly lines. Linker regions (are short sequences that connect adjacent domains or modules) were also included along

with A-PCP-C domains while swapping the modules, which resulted in the efficient production of novel peptides (Bozhüyük *et al.*, 2018).

In the following section, an explanation is given for how small molecules are represented in computational chemistry methods, which will be useful when structural comparisons of fungal A-domain substrates and NRP monomers was done. Ideally, molecules should be encoded into a format without losing too much information about the structure and stereochemistry. Also, to handle a collection of molecular structures, they must be encoded into a simple format, such that they could be easily queried from a database and compared.

1.9 Encoding of small molecules

1.9.1 Molecular representation

There are many ways of encoding small molecules in a machine-readable format (Engel 2003). One dimensional Simplified molecular-input line-entry system (SMILES) is a text-based molecular representation format with ASCII characters for atom and bonds, e.g., SMILES for Propyl alcohol is CCCO. Depending on which heavy atom is used to start writing SMILES strings, multiple SMILES strings could be produced, which introduces ambiguity. Canonical SMILES were developed to overcome this problem, but they also produce redundant entries for some molecules. SMARTS can describe small fragments of molecules. SMARTS has additional symbols than SMILES that describe special atomic properties, e.g. valence, ring connectivity, chirality, etc. SMARTS strings give flexibility in molecular substructure comparisons; hence they are used to query large molecular databases. SMILES are a subset of SMARTS representations. To provide a unique and non-proprietary standard identifier, which might circumvent the problems associated with other identifiers, International Chemical Identifier (InChI) was introduced in 2006 (McNaught 2006). InChI has a hierarchical layered format to represent different features of a chemical structure (Heller *et al.*, 2015). These different layers could encode information such as charge, stereochemistry, empirical formula, atom connectivity and number and connectivity of hydrogen atoms.

1.9.2 Molecular descriptors

Molecular descriptors are numerical representations of molecular characteristics derived from their structures or chemical constitutions. These descriptors are calculated either by an algorithm through a predefined set of rules or by experimental measurements or quantum mechanical calculations.

1D or 0D descriptors are calculated from the empirical formula of the molecules. These are simple descriptors such as the number of atoms, bond order, molecular weights. 2D descriptors are computed using 2D structures of molecules; examples include topological distances and charges, etc. Similarly, 3D descriptors need to be derived from the 3D structural representation of molecules, such as surface/volume ratio, solvent accessible surface area. 4D, 5D and 6D descriptors are composite descriptors, they encode dynamic information of the molecular structure along with other molecular properties. Dragon 7.0, commercial software can compute as many as 5270 molecular descriptors. There are other freeware alternatives for molecular descriptor computations such as MODEL (3778 descriptors, Li *et al.*, 2007), PaDEL (1875 descriptors, Chun 2010).

1.9.3 Molecular fingerprints

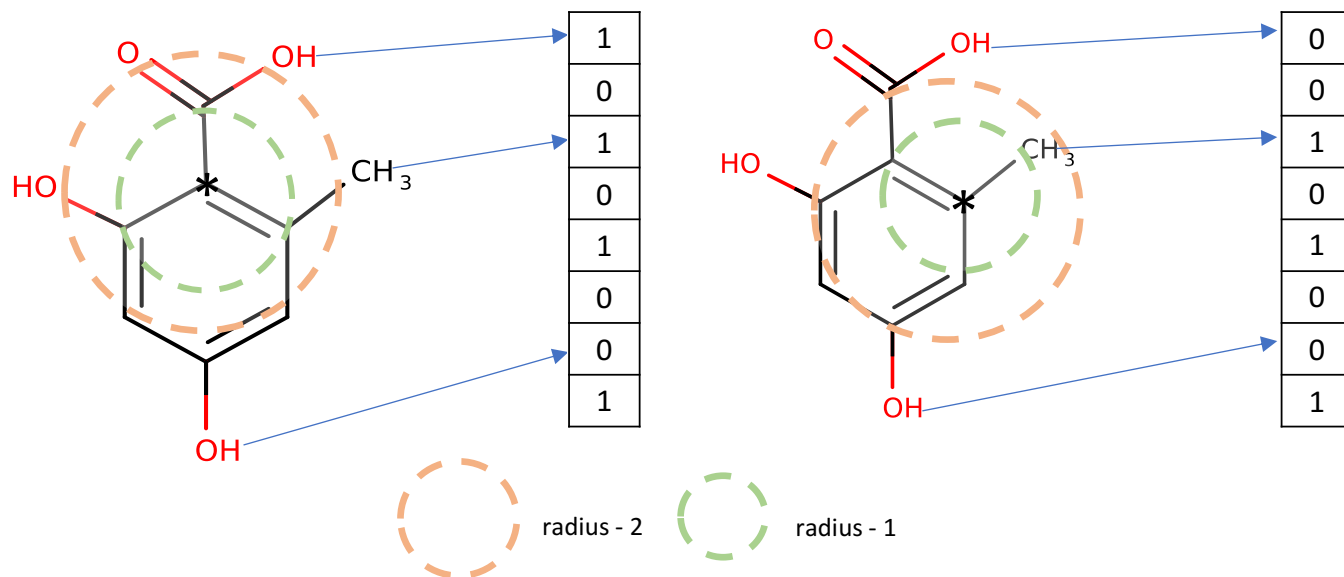


Figure 4. Schematic representation for Morgan fingerprint computation. For each heavy atom, all the neighborhood atoms with a given radius (radius signifies the number of atoms to traverse around the atom) are found. The circular neighborhood for two heavy atoms (shown with an asterisk) with radius 1 or 2 (circles with dotted lines) are shown. After finding these sets of atoms or substructures,

they are encoded into a bit vector of a fixed length. Bit value indicates the presence or absence of a substructure, which is shown by an arrow pointing at their positions in the bit vector.

Molecular fingerprints allow encoding of molecules into bit vectors. Bits map to either presence (or absence) or counts of molecular properties and substructures. Simple fingerprints include atom pair fingerprints, atom connectivity information for pair of atoms is stored. Molecular substructure-based fingerprints are Molecular ACCess System (MACCS) keys, which are binary fingerprint bit vectors with the size equal to 166, each of the bit encodes predefined MACCS keys. Each of the MACCS keys is predefined substructures and atom types. As substructure types are limited, this imposes difficulty in encoding diverse chemical compounds. Morgan fingerprints, which are widely used fingerprints that calculate circular neighborhood around heavy atoms dynamically hence could be used for chemically distinct molecules.

Morgan fingerprints are computed as shown in Fig 4. Generally, functional-class fingerprints (FCFP) definitions for Morgan fingerprints are used to build structure-function relationships. In FCFP, functionally similar atoms could be used interchangeably e.g., OH and SH groups are treated same since both O and S atoms act as hydrogen bond donors. CDK (Willighagen *et al.* 2017), RDKit (Landrum 2006) and Open Babel (Boyle *et al.* 2011) are widely used packages to handle small molecules into various fingerprints and other structural analysis.

1.9.4 Molecular similarity

Tanimoto (or Jaccard) and Dice are widely used coefficients for molecular structure comparisons. Other distance-based metrics include Euclidean and Manhattan distances. A comparison of eight metrics by Bajusz *et al.*, 2015, has put Tanimoto, Dice, Cosine and Soergel coefficient in the same basket as these produce similar rankings for structurally similar compounds.

Tanimoto and Dice coefficients are calculated as shown in below.

$$Tanimoto(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

$$Dice(A, B) = \frac{2 * |A \cap B|}{|A| + |B|}$$

A and B are sets of molecular descriptors or properties of two molecules

Distances are calculated as 1- Tanimoto (A, B) or 1- Dice (A, B).

1.10 Computational methods for A-domain substrate specificity predictions

For many bacterial A-domains, substrate specificity determinants (NRPS code) have been determined through experiments, although in absence of homologs (with known substrates) it is difficult to assign specificity for novel fungal A-domains. Computational methods are developed to leverage this specificity information for accurate substrate predictions.

A-domain substrate specificity prediction tools that are developed so far (Table 3) use protein sequences or features derived thereof for the prediction of the substrates. Although single residue changes have a detrimental effect on non-covalent interactions, shape and charge microenvironment, hence structural insights are crucial for successful specificity prediction. For other protein classes, 3D structure-based ligand-protein fragment-fragment interactions have been exploited to predict possible ligands with the use of artificial neural network models (Tang *et al.* 2014). Such data-driven approaches are feasible when enough ligand-protein complex structural information is available. The scarcity of experimentally elucidated structural data for fungal A-domains has hindered the development of structure-based machine learning methods. More efforts shall be devoted to crystallizing A-domain structures to decipher substrate recognition mechanisms.

The tools listed in Table 3 use more bacterial sequences for training classifier, hence they work well with bacterial sequences but fail to give accurate predictions for fungi. For example, SANDPUMA uses only 9.6% (90) fungal sequences to train its algorithms in contrast to 88.3% (819) bacterial sequences. This shows bias towards studying prokaryotic over eukaryotic NRPS mega synthetases.

1.10.1 SANDPUMA: an ensemble classifier

SANDPUMA (Chevrette *et al.*, 2017) is a relatively recently developed tool for substrate specificity inference. This tool is incorporated into a widely used pipeline, antiSMASH (Blin *et al.*, 2017), which is used worldwide for secondary metabolite structure prediction by analyzing genomic sequences. SANDPUMA is an ensemble classifier that combines binding site (active site motif and support vector machine) and full-length sequence (profile HMM and prediCAT) based methods to

predict substrates of A-domains. SANDPUMA considers percent sequence identity (with training examples) and predictions from individual algorithms to assign specificity to A-domains.

Table 3: A-domain substrate specificity prediction tools. Accuracy and F measure values are listed here (obtained from the publication) for each algorithm with their datasets. Fungal A-domain substrate specificity accuracy or F measure is mentioned for NRPSpredictor2 and LSI. SB = Substrates, SQ = Sequences, F= F measure.

Tool	Algorithm	Dataset	Accuracy	Current Status
SANDPUMA (Chevrette <i>et al.</i> , 2017)	Decision tree	928 SQ 104 SB	0.84	incorporated into antiSMASH 4.0
Virtual screening (Lee <i>et al.</i> , 2015)	Ligand docking	10 (structures) 12 (models) 59 SB	1.0 (structures) 0.61 (models)	no web tool
SQL-NRPS (Knudsen <i>et al.</i> , 2015)	Sequence learner	537 SQ 37 SB	0.71	active
NRPSpredictor2 (Röttig <i>et al.</i> , 2011)	Support vector machine	576 SQ 75 SB	F 0.94 (bacterial) F 0.84 (fungal)	inactive
LSI (Baranašić <i>et al.</i> , 2014)	Latent semantic indexing	397 SQ 47 SB	0.89 (bacterial) 0.85 (fungal)	inactive
NRPSsp (Prieto <i>et al.</i> , 2012)	Hidden Markov models	1578 SQ	0.86	active
NRPS/PKS substrate predictor (Khayatt <i>et al.</i> , 2013)	Hidden Markov models	571 SQ 58 SB	0.66	inactive

1.10.2 Limitations and drawbacks of the previously developed tools

As already mentioned above, the main drawback of the existing tools is the strong bias towards bacterial sequences, hence resulting in unreliable predictions for eukaryotic A-domains. The tools shown in Table 3 do not give accurate substrate predictions for fungal A-domains, e.g. A-domains which do not share NRPS code similarity with training set dataset sequences. This could be partly because of the usage of sequence-derived features for specificity prediction, which might not accurately model atomic-level interactions between substrates and NRPS code residues.

SANDPUMA is incorporated into antiSMASH, hence it can be used through antiSMASH webserver (or locally by downloading binary files). Other tools, which have a web-based interface and are currently active include SEQL-NRPS and NRPSsp and remaining tools are either inactive or unavailable for use through a web interface.

As stated before, the primary aim of this research work was to develop a fungal adenylation domain substrate specificity prediction tool, this was accomplished by training a neural network model. These models belong to widely used machine learning methods and hence a short overview of machine learning and neural network methods is given in the following sections.

1.11 Machine learning

Machine learning (ML) is defined by Stanford University as “Science of getting computers to act without being explicitly programmed” (<http://cs229.stanford.edu>, CS229: Machine Learning). It is based on a premise that computers could learn the inherent pattern in the data so that derived models are useful in the future. ML could be used in scenarios, where humans cannot discern patterns in high dimensional complex data. Depending on the problem at hand and the amount or type of data one possesses, a pertinent algorithm needs to be selected. While building ML models, a few things that need to be carefully considered are the correct splitting of data for the training (and the validation and testing purposes), the proper handling of hidden variables or experimental noise and defining the clear objectives for the prediction model (Rile 2019).

We have witnessed advances in ML in image recognition (Simonyan *et al.*, 2014), spam detection (Crawford *et al.*, 2015), and sentiment analysis areas (Gautam *et al.*, 2014). ML algorithms have also been successfully used in biological applications, e.g., protein structure prediction, protein-ligand

fragment binding predictions (Tang *et al.*, 2014), protein folding, virtual screening, etc. (Carpenter *et al.*, 2018). Recently held critical assessment of protein structure prediction (CASP13) competition (to test the ability of computational methods to predict 3D structures of proteins from sequences) was won by Google's DeepMind firm with their AlphaFold algorithm (AlQuraishi *et al.*, 2019). AlphaFold uses deep learning to predict distances between protein atoms by training on Protein Data Bank (PDB) three-dimensional structural data. Such advances surely give hope that ML could be wisely used to solve long-standing, challenging biological problems.

Principally, ML methods could be subdivided into two types, supervised and unsupervised methods. For supervised ML, ground truth labels are required i.e. the mapping between input and output needs to be known beforehand, though, labeling of samples is a time consuming and labor-intensive process. Support vector machines, random forest, k-nearest neighbor classifications are examples of supervised ML algorithms. Classification algorithms would categorize samples to a predefined set of groups or classes e.g. logistic regression. These could be binary or multi-class classification, where two or multiple classes could be predicted respectively. Regression methods involve predicting continuous value as an output, e.g. linear or polynomial regression methods.

No verified labels or predefined classes are required for unsupervised methods. This approach could be used for certain tasks such as clustering (k-means clustering) or dimensionality reduction (principal component analysis) analysis. Semi-supervised learning methods combine a higher amount of unlabeled data with generally limited labeled data to leverage the gap between two data sources.

1.11.1 Artificial neural networks

Artificial neural network (NN) is a supervised learning algorithm, proposed long back in the 1940s, which was inspired by the human brain's neuronal structure. Neuropsychologist Donald Hebb had proclaimed that "Neurons that fire together, wire together" to illustrate neuronal communication through neurotransmitter release, which forms the basis for artificial NN algorithms. In the human brain, neurons are heavily interconnected to each other and are responsible for the transmission of information. Likewise, neurons are a fundamental unit of artificial NN. Neurons are mathematical functions, each of them is associated with certain weights and biases. Artificial NN architecture is

such that neurons are arranged in multiple layers and are interconnected to each other. Simplified representation for an artificial NN with 3 layers is shown in figure 5.

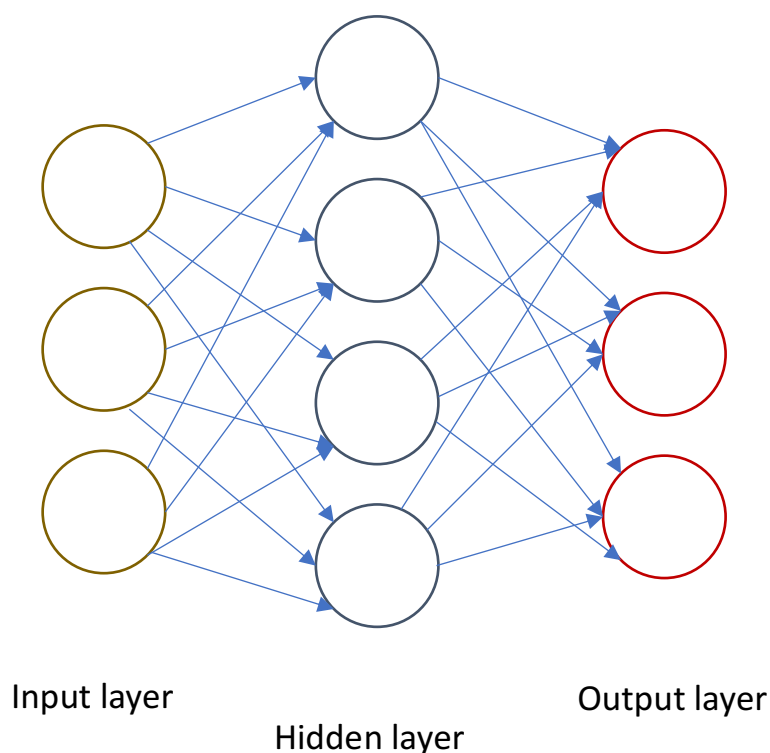


Figure 5. Simple artificial NN with three layers. An input layer, to which input is supplied and hidden layer where all the computations happen and all neurons are connected between this and other layers, and the output layer gives the encoded outputs. Connections between neurons of different layers are shown with arrows (blue).

There are certain NN parameters that need to be optimized to map input features to outputs correctly. Parameters for neural network model developed here (for A-domain substrate specificity prediction) were optimized and hence they are mentioned below briefly.

Activation or transfer function is a nonlinear complex function that maps input features to the output. Examples include binary step, sigmoid, Rectified Linear Unit (ReLU), Scaled Exponential Linear Unit (SELU) functions, etc. The input layer is the first layer of neurons to which input data (or variables) is supplied. Hidden layers are the workhorses of the NN, where all the computations happen and all the neurons in this layer are connected to every other neuron in the next layer. The

output layer is the last layer, whose neurons are encoded as the possible outputs (finite) expected from the model. Loss or cost function determines the error rate of NN predictions by comparing predicted and expected outputs. The learning rate, which is also called step size, is a scalar value between 0.0 to 1.0 that determines how fast the NN model learns weights. The batch size determines the number of samples passed to through NN, after dividing training data into different partitions. One epoch is completed when a complete dataset is presented to NN once.

Weights determine the connectedness of neurons between two layers and their magnitude signifies the strengths of these connections. Bias is incorporated to ensure that a neuron is activated. In the case of artificial NN, “learning” refers to finding the right combination of weights and biases, such that output is precisely predicted.

Artificial NN topology, the way neurons are interconnected affects how the predictions are made. The inclusion of more layers adds more complexity and hence the time required to train the algorithm. As for any ML approach, the training dataset should be a large and well-represented sample (of actual phenomena under investigation), as this determines generalizability and usability.

1.11.2 Evaluation of predictor performance

After the ML model has been built, its performance needs to be evaluated on previously unseen data to assess its generalizability. It would also be useful to know if the predicted output correctly matches with an expected output. In the case of classification, one could check if the model assigns class membership correctly. The simplest measure to determine this is to calculate accuracy. For binary classification, with two classes A and B, there are four scenarios (TP, FP, FN, and TN) as shown in Table 4.

Table 4. Confusion matrix, for a binary classification example, four outcomes could be expected.

	Class A (predicted)	Class B (predicted)
Class A (actual)	TP (true positive)	FN (false negative)
Class B (actual)	FP (false positive)	TN (true negative)

The accuracy of the model is calculated by taking proportions of values obtained in the confusion matrix as shown in the equation below.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

The other measure that is widely used is the F1 Score, which is defined as a harmonic mean of precision and recall. Precision, recall and F1 score are calculated as follows. Precision is determined by counting, of all predicted memberships to a class, how many belong to it. Recall measures, of all class members (belonging to a one class), how many members are correctly predicted by the model to belong.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1Score = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

Accuracy and F1 scores are insensitive to the imbalanced nature of the dataset, hence in such scenarios, alternative measures such as the Matthews correlation coefficient (MCC) is used.

MCC is a balanced measure to assess the performance of the model. It varies from -1 to 1, where 1 represents the complete agreement between prediction and expected output and -1 show perfect disagreement. Zero suggests no better prediction than a random guess.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}}$$

As all the topics that relate to the research work done here are introduced, paragraphs below briefly mention the aims of the thesis and composition of each of the sub-sections. The primary aim of the thesis was to develop a fungal A-domain substrate specificity prediction classifier. Neural network-based substrate specificity classifier (NNassc) was trained by combining the physicochemical and

structural properties of NRPS code residues. The data used to train the model was solely from fungi, hence this model shall be specifically used to assign substrates for fungal A- domains.

“Methods” section gives information about A-domain substrate specificity data curation, from the published literature and subsequent phylogenetic analysis. Also, the neural network model hyperparameter optimizations and development workflow for substrate prediction is explained in detail. To evaluate the developed model, internal and external validation datasets were used and predictions for each of these datasets are listed. In the “Results and discussion” section, investigations into substrate specificity differences and phylogenetic relationships between fungal and bacterial A-domains (and NRPS codes) are included. Clustering of full-length A-domain sequences captures overall differences in different NRPS mega synthetase and product type, while NRPS code residues give an idea about substrates activated. Comparisons of SANDPUMA (an ensemble classifier for adenylation domain substrate specificity predictor) with neural network model (developed here) using independent validation datasets are incorporated. Substrates for which NNasc gives correct predictions are listed and strategies for the improvement of prediction performance are also mentioned.

2. Methods

2.1 A-domain substrate specificity data

The collection of A-domain sequences used in this work combined data from two sources:

i) Existing mixed (eukaryotes and bacteria) collection: NRPSpredictor2 is a support vector machine-based A-domain substrate specificity classifier which also contains fungal specific predictor (Röttig *et al.*, 2011). NRPSpredictor2 dataset of A-domains with experimentally verified substrates consisted of 576 sequences from bacteria and fungi. This collection contains 111 fungal sequences.

ii) New fungal collection: Manually curated list of experimentally characterized fungal A-domain sequences published during the period 2011-2018. We assumed that all sequences published before 2011 have been included in the NRPSpredictor2 dataset; therefore, the new collection aimed at the sequences published thereafter. The sequences were collected by text mining of PubMed and Google Scholar, searching for terms related to fungal A-domain substrate specificity. It was carefully checked that A-domain substrates are elucidated by ATP-PPi exchange assay; computationally predicted substrate specificity information was not included in the dataset. For the identified domains, corresponding UniProt identifiers were obtained from the UniProtKB database. The new fungal collection counted 33 sequences.

The final dataset (merged from i and ii) referred in the following sections as fungal A-domain dataset (Appendix, Table 1) is populated with 144 fungal A-domain sequences with characterized substrates.

2.2 Phylogenetic analysis

2.2.1 Phylogenetic analysis of fungal A-domain sequences

The multiple sequence alignment was obtained for 144 sequences of the fungal A-domain dataset by using MUSCLE (Edgar 2004). The maximum likelihood phylogenetic tree was built by PhyML 3.0 with 100 bootstrap cycles. Bacterial phenylalanine binding A-domain (PheA, UniProt identifier: P0C061) was used as an outgroup. The phylogenetic tree was visualized and processed further using FigTree v1.4.4. Leaves of the phylogenetic tree were labeled with substrate specificity information for each A-domain.

2.2.2 Extraction of C-domains and phylogenetic analysis

For every A-domain with known substrate specificity, C-domains upstream to it were extracted from complete NRPS mega-synthetase protein sequences. These pairs of domains (Upstream-Downstream) were obtained by matching with Pfam models (for A-domains: PF00501 and C-domains: PF00668). Coordinates for domain boundaries were obtained from the Pfam database, which stores domain architecture for protein families. 113 out of 144 A-domains (fungal A-domain dataset) had C-domains upstream as suggested by Pfam domain matches. After extracting 113 pairs of domains, substrate specificity information from A-domain was transferred to the corresponding upstream C-domain. Truncated or incomplete C-domains were not included in multiple sequence alignment and phylogeny analysis. Eventually, 94 C-domains were used to build a rooted phylogenetic tree using the same procedure as described for A-domains. PheA C-domain was used as an outgroup and leaves were labeled by transferring substrate specificity information from downstream A-domain.

2.3 Clustering of bacterial and fungal A-domain NRPS codes

NRPS codes were obtained for all sequences from the fungal A-domain dataset. After removing redundant NRPS codes that bind the same substrates, 262 codes were retained. Bacterial NRPS codes were directly used from the NRSPredictor2 dataset, without updating their dataset. Amino acid residues of the code were encoded by physicochemical properties essential for binding to small molecules: size (WOLS870102), hydrophobicity (WOLS870101) and electronic properties (WOLS870103) (Sandberg *et al.*, 1988). The last amino acid of the code, lysine, was not included, as this catalytic residue is perfectly conserved in all NRPS codes. Each amino acid of the code has three associated properties; hence descriptor vector was created with a total of 27 properties (for 9 NRPS code amino acids). Euclidean distances were computed for these descriptor vectors and were clustered using Pvcust (Suzuki *et al.*, 2015) package in R.

2.4 Clustering of non-ribosomal peptide monomers

Non-ribosomal peptide (NRP) monomers and A-domain substrates from different NRPs are deposited in the Norine database (Caboche *et al.*, 2008). 13 new substrates were obtained from the fungal A-domain dataset, these substrates are not found in the Norine database. Monomer simplified molecular-input line-entry system (SMILES) strings were obtained from the PubChem database if

those were absent in the Norine database. In total, 204 (191 from Norine and 13 new) SMILES strings were used for clustering. SMILES were then transformed into Morgan fingerprints using the RDKit package (Landrum, 2016). Morgan fingerprints were used with a radius of two (around each heavy atom). Tanimoto coefficient values were computed for each pair of monomer fingerprints, and distances (1- similarity) were used to construct dendrogram using hclust package in R.

2.5 Development of NN-based A-domain substrate specificity classifier (NNassc)

NNassc was programmed in Python scripting language and RDKit package (Landrum, 2016) and Keras (Chollet *et al.*, 2015) with TensorFlow (Abadi *et al.*, 2016) backend was used for the encoding of substrates and training a neural network respectively.

2.5.1 Dataset preparation for NNassc

NRPS code and substrate information were obtained for experimentally characterized A-domains (see 3.1) from the research articles describing them. Substrate SMILES were procured from the PubChem database. There were 41 different substrates in the fungal specificity prediction dataset. After removing redundant NRPS codes (two or more instances of same NRPS code-substrate pair), the final dataset was reduced to 136 NRPS codes. Substrates along with number of NRPS codes that were used for training NNassc are listed in Table 5.

Table 5. NNassc substrate specificity prediction dataset. Substrates and the number of NRPS codes identified for that substrate through experimental characterization are listed. In the last row, substrates with only one NRPS code per substrate are listed.

Substrate	# NRPS codes	Substrate	# NRPS codes
Alanine	14	Tyrosine	4
Valine	11	Serine	4
Leucine	10	Isoleucine	3
α -amino-iso-butyric acid	10	Glycine	3

Tryptophan	9	Homoserine	3
α -amino-adipic acid	9	Anthranilic acid	3
Phenylalanine	8	Pipecolic acid	2
Isovaline	7	Glutamine	2
Proline	6	Cysteine	2
Tyrosine	4	β -alanine	2
Fumaric acid	2	4-hydroxy-phenylpyruvic acid	2

s-nmethoxy-tryptophan, ornithine, meval, mephe, hmp-D, hiv-D, hiv, glutamic acid, aoda, D-alanine, aeo, phenylpyruvic acid, meOhVal, Indole-3-pyruvic acid, Grifolic acid, cisAMHO, α -hydroxy-isocaproic acid, 5-methyl-orsellinic acid, 4-hydroxy-ornithine

cisAMHO : N(δ)-cis-anhydromevalonyl-N(δ)-hydroxy-L-ornithine, D-ala : D-alanine, abu : α -amino-isobutyric acid, aad : α -amino adipic acid, pip : pipecolic acid, hiv : 2-hydroxy isovaleric acid (hiv-d, D isomer), bmt : (4R)-4-[(E)-2-butyl]-4-methyl-L-Thr, aoda : (S)-2-amino-8-oxodecanoic acid, meval : N-methyl-valine, mephe : N-methyl phenylalanine, hmp-D : D-2-hydroxy-3-methylpentanoic acid, aeo : L-2-amino-8-oxo-9,10-decanoate.

2.5.2 Preprocessing

2.5.2.1 Encoding of NRPS code residues

NRPS codes were encoded by two sets of features, physicochemical & molecular structural properties. The physicochemical properties are procured from Sandberg *et al.* 1998 (Wold encoding), those are the size (WOLS870102), hydrophobicity (WOLS870101) and electronic properties (WOLS870103). NRPSpredictor2 (Röttig *et al.*, 2011) also uses Wold encoding to encode NRPS codes for A-domain substrate prediction. Structural features were obtained by encoding each of NRPS code amino acids into Morgan fingerprint (see fig 4, for computation of fingerprints) bit vector (size = 4096) with functional class fingerprints definitions. Bit positions that are ON (i.e., the substructure is present) in at least one amino acid (among all the NRPS code amino acids) were considered while others were discarded. Out of 4096 bits in amino acid fingerprints, only 30 bits were ON in at least one of the amino acids found in the code and thus retained. A new bit vector of

size 30 was created. Eventually, each amino acid of the code is represented by 3 physicochemical and 30 structural properties derived from Wold encoding and Morgan fingerprints respectively. With 9 amino acids of the code, the input feature vector size was 297 ($9 * 3 + 9 * 30$).

2.5.2.2. Encoding of substrates

Previously developed tools for A-domain substrate specificity prediction did not consider substrate structure while training. Here, substrate SMILES were encoded in Morgan fingerprint bit-vectors and classifier output was a bit vector (size = 1024) as opposed to simple substrate class in previous tools. Each bit of the bit vector signifies chemical features or substructures the substrate possesses.

2.5.3 Parameter optimization and training NNasse

Neural network hyperparameters were optimized with Hyperopt (Bergstra J. *et al.*, 2013), a python library that allows optimization with real-valued, discrete and conditional dimensions. The number of input features equal to the size of the feature vector (size = 297) was specified by *the input_dim* parameter in the Keras library. *input_dim* parameter was kept unchanged during the optimization procedure.

Four parameters: number of neurons in the first hidden layer, type of optimizer, batch size and number of epochs, were optimized through Hyperopt by providing initial search spaces. The search spaces were arbitrarily chosen and could be different depending on the training dataset and the type of features used. Parameters with corresponding search spaces are shown below.

- Number of neurons in the first hidden layer

$X = \text{integer}(\text{size of input feature vector}/d)$

where $d = \{20, 10, 5, 2, 1, \frac{1}{2}, \frac{1}{5}, \frac{1}{10}, \frac{1}{20}\}$ and size of input feature vector = 297

- Batch size = {10, 20, 30, 40, 50}
- Epochs = {50, 75, 100, 125}
- Optimizer = {'adam', 'RMSprop', 'sgd'}

The number of neurons in the output layer was equal to the size of the substrate Morgan fingerprint bit vector (bit size = 1024). Sigmoid activation function was used for the output layer. Unlike SoftMax function, Sigmoid function allows treating individual class probabilities independent without squashing their sum equal to 1. As a result of this, each bit position of bit vector could be predicted independently. The output layer produces probability values suggesting confidence for the presence or absence of a substructure. As the output layer encodes different substructures for the substrate bit vector, this is a multi-label classification example. Binary cross-entropy loss was used, which calculates the loss for each of the multiple labels separately.

2.5.4 Matching predicted bit vector with substrates from training dataset

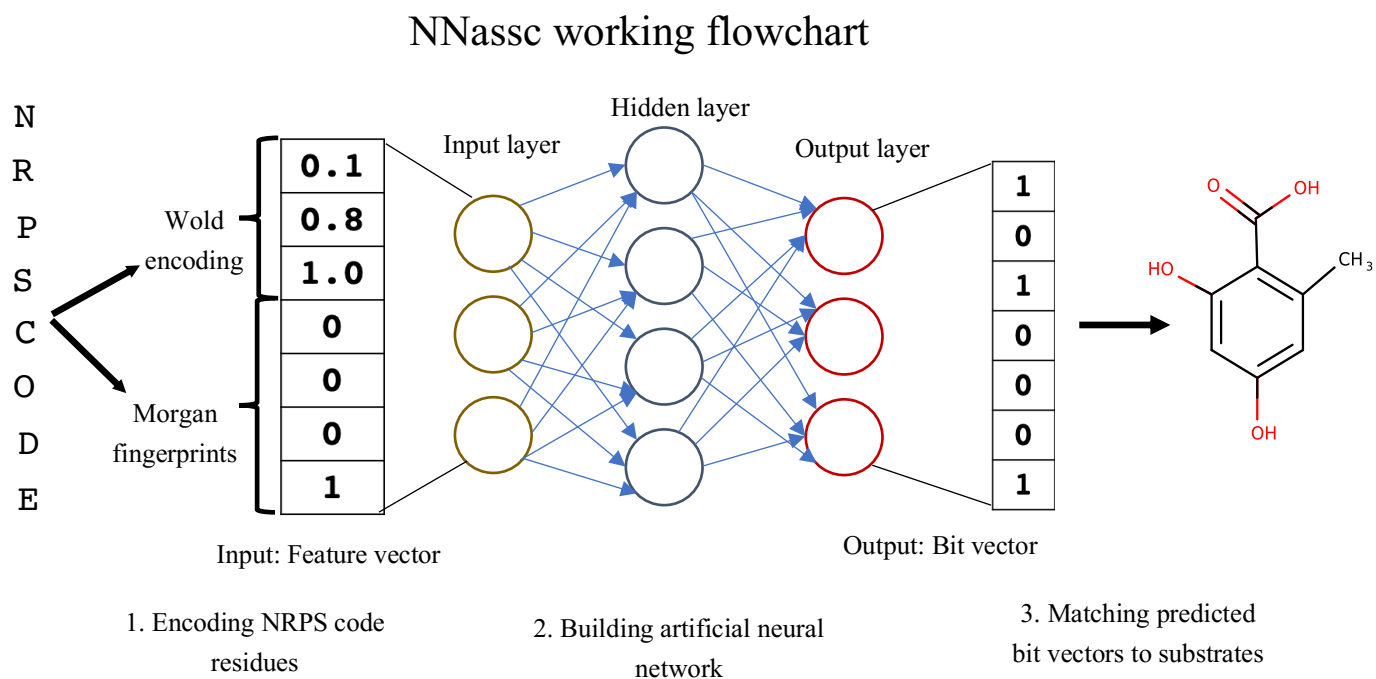


Figure 6. Overview of NNassc method. NRPS code residues are encoded into a feature vector and are provided as an input for the ANN. The input is processed through ANN to produce a 1024 size bit vector which gives a probability for the presence or absence of substructures. The output bit vector is matched with possible substrates by comparing these substructures.

After predicting the fingerprint bit-vector probabilities for each bit position, it was important to know what substrate it shares close similarity with, to assign the specificity for NRPS code. To transform these probability values into a binary encoding, a generally accepted cutoff of 0.5 was used. Bit

positions with probability values above 0.5 were set to 1 (ON or substructure is present) otherwise 0 (OFF or substructure is absent). Such a bit-vector representation has been used previously to assign metabolites to a mass spectrum (Dührkop *et al.*, 2015). In the end, the Tanimoto similarity coefficient was used to compare predicted bit vector with training dataset substrate fingerprint bit vectors. Rank ordered list of possible substrates was obtained with the decreasing values of Tanimoto similarity. Top three or five predictions were checked to see if predicted bit-vectors match the actual substrates.

20 different models were generated by dividing A-domain substrate specificity data randomly into training (80%) and validation dataset (20%). The best model was chosen by assessing accuracy on the internal and external validation dataset. As there are many model selection strategies, different ways of selecting models could be tested in the future.

2.5.5 Validation and benchmarking

2.5.5.1 Internal validation dataset

For many substrates, only a few NRPS codes were identified so far, hence dividing them into separate training, validation and testing dataset was not feasible. To find the structurally similar substrates and combine them into classes for validation, this clustering exercise was done. Substrates from the training dataset were encoded into Morgan fingerprints and were clustered by their structural similarity. This clustering helped find 3 classes (Table 6) of substrates from the training dataset substrates. After dividing substrates (and corresponding NRPS codes) into 3 classes, 20% dataset was used for validation, which comprised of representatives from each of these classes.

2.5.5.2 External validation dataset

To assess the performance of NNassc and demonstrate its generalizability on novel data, I used two sets of NRPS codes that were not included the training dataset. The first source of these NRPS codes was from an article that describes cycloaspeptide biosynthesis (Mattos-Shipley *et al.*, 2018) and the second set was from our collaborators (Dr. Daniel Berry, Massey University, New Zealand) obtained through personal communications. In total, there were 10 NRPS codes in the external validation dataset. The prediction was run by the NNassc and SANDPUMA. The top three substrate predictions obtained through NNass for each NRPS code are shown in Table 11.

Table 6. Substrate classes obtained by structural similarity-based clustering (for validation of the NNasse). Representatives from each substrate class and corresponding NRPS codes were included in the internal validation dataset.

Class 1	Class 2	Class 3
Valine	Tryptophan	Ornithine
Alanine	Indole-3-pyruvic acid	Glutamine
Allo-isoleucine	s-n-methoxy tryptophan	2-amino adipic acid
2-aminobutyric acid	Fumaric acid	Homoarginine
Leucine	Phenylpyruvic acid	Arginine
D-2-hydroxyisovalerate	4-hydroxy phenyl pyruvic acid	2S-2-amino-8-oxodecanoic acid
α -hydroxy-isocaproic acid	Phenylalanine	2-amino-8-oxiran-2-yl-8-oxo octanoic acid
D-2-hydroxy-3-methyl-pentanoic acid	Tyrosine	N(δ)-cis-anhydromevalonyl-N(δ)-hydroxy-L-ornithine
Beta-alanine	Grifolic acid	
Glycine	5-methyl-orsellinic acid	
Pipecolic acid	Anthranilic acid	
Proline		
Serine		
Homoserine		
Cysteine		

3. Results and discussion

3.1 Phylogenetic analysis of fungal A-domains

In the following subsections, I shall describe fungal A-domain (full length) clustering depicted in Fig 7. To study phylogenetic relationships between fungal A-domains and to better understand the evolution of their substrate specificity this analysis was done. This helped decipher relationships between various fungal A-domains that are either a part of the same NRPS or bind the same or similar substrate. Clustering results for different NRPS mega synthetases are summarized in Table 7 and subsections are given below.

3.1.1 A-domains activating uncommon non-proteogenic substrates

ArmA A-domain (G3FLZ5), experimental studies (Misiak *et al.*, 2011) suggest probable substrates leu, thr or val as, but it is clustered with non-proteogenic substrate 5-methyl-orsellinic acid and grifolic acid binding domains.

3.1.2 ACVS synthetase

ACVS synthetases (Q9C1G0, P25464, P26046, P19787, and P7742) are involved in L- δ -(α -amino adipoyl)-L-cysteinyl-D-valine (ACV) biosynthesis, which is a penicillin and cephalosporin precursor. ACVS synthetases belong to a gene cluster horizontally transferred from bacteria (Brakhage 1998). All the A-domains belonging to ACVS synthetases tend to cluster together in a single clade and do not cluster with other A-domains binding the same substrates, for example, alpha-amino adipic acid binding A-domains from ACVS synthetases and lysine biosynthesis (M2PFR6, P007702, P40976) do not cluster together. Lysine biosynthesis A-domains belong to NRPS-like carboxylic acid reductases family and they utilize ATP or NADPH for the reduction of carboxylic acids.

3.1.3 PKS-NRPS hybrids

A-domains from hybrid PKS-NRPS synthetases A0JJU1 (Tenellin Synthetase from *Beauveria bassiana*), A1CLY8 (Mycotoxin biosynthesis, from *Aspergillus clavatus*), A8KNE2 (from *Penicillium expansum*), B6F209 (Cyclopiazoic acid synthetase from *Aspergillus oryzae*), C9K4U2

(from *Aspergillus flavus*), Q0D159 (Isoflavipucine biosynthesis, from *Aspergillus terreus*), Q5ATG8 (Aspyridones biosynthesis, from *Aspergillus nidulans*), Q3L7Y0 and C6KDY5 belong to Ascomycota division of fungi and are all clustered together in a single monophyletic group. Homoserine and aromatic amino acid (tyr, trp and phe) binding A-domains form a separate clade, However, the clade for aromatic acid binding A-domains is interspersed by two smaller substrates (ser or leu) binding domains.

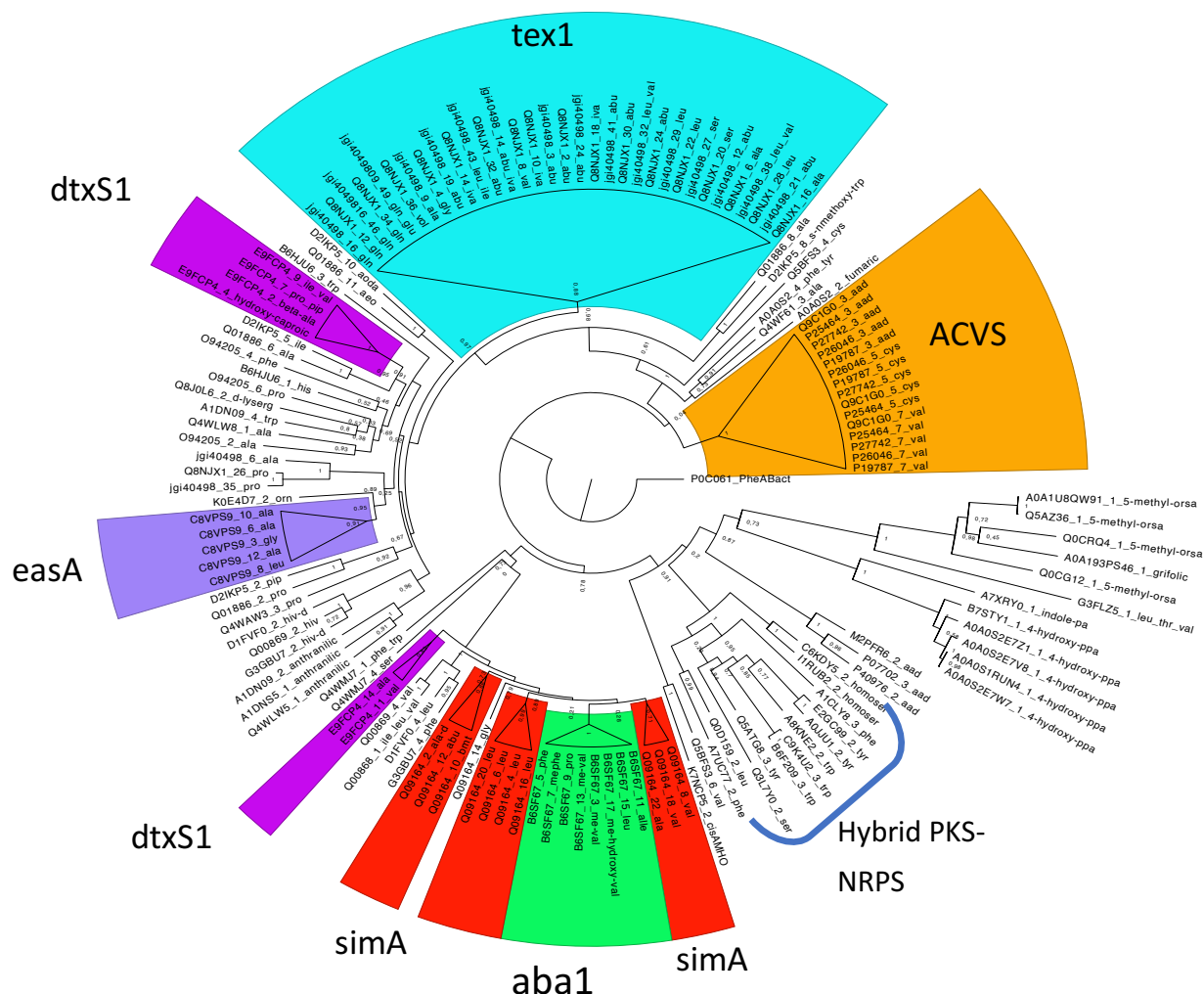


Figure 7. Fungal A-domain (full-length) phylogenetic analysis. Leaves of the tree were labeled with substrate specificity and clades for A-domains from the same NRPS are colored and the name of the gene is also mentioned. UniProt identifier and A-domain number as a part of NRPS protein was

obtained from the Pfam database (only A and C-domains were numbered in a complete NRPS mega-synthetase protein sequence).

Table 7. Fungal A-domain (full length) phylogeny analysis. NRPS mega synthetase type (UniProt identifier) and their clustering behavior are listed. For A-domains that cluster together, corresponding substrates are mentioned along with some information about clusters (number of clades for A-domains from the same NRPS type or substrate specificity-wise clustering, etc.).

NRPS	Substrates	A-domain clustering and NRPS architecture
ACVS synthetase	aad, cys, val	cluster together in a single clade
Siderophores (Q5BFS3 and K7NCP5)	val cisAMHO	cluster together irrespective of binding different substrates
Emericellamides (C8VPS9)	ala, gly, leu	cluster in a single clade
Aureobasidin A1 biosynthetic complex (B6SF67)	phe, methyl-phe, pro, methyl-val, methyl-hydroxy val, leu, allo-isoleucine	cluster in a single clade
Desipeptides	hiv, hiv-d	Two sets of clades
D1FVF0, G3GBU7, Q00869	val, leu, phe	
HC-toxin (D2IKP5 and Q01886)	s-nmethoxy trp, aoda, ile, pip	All A-domains of this NRPS cluster separately. Pro and aoda binding A-domains cluster with domains with similar substrates

Destruxins (insecticidal cyclic hexadepsipeptides) E9FCP4	ile or val, pro or pip, beta-alanine, 4-hydroxy-caproic acid ala, val	A-domains of this NRPS cluster in two clades with 2 set of substrates as shown
Cyclosporine synthetase (Q09164)	val, ala D-ala, abu, bmt leu	Three clades with the substrates shown
Peptaibols (Q8NjX1), jgi40498	Binds various small aliphatic and polar substrates pro	For Q8NjX1 and jgi40498, A-domains are clustered in a single clade except for few pro and ala
A1DN09, A1DNS5, Q4WLW5	anthranilic acid	A-domains from different NRPSs binding this substrate cluster together
Uncommon substrates	4-hydroxy-phenylpyruvic acid, phenylpyruvic acid, indole-3-pyruvic acid 5-methyl orsellinic acid, grifolic acid leu or thr or val	A-T-TE domain architecture and lack C-domain part of NRPS like enzymes. A-domain followed by NAD binding domain. ArmA is a part of A-T-TE

cisAMHO - N(δ)-cis-anhydromevalonyl-N(δ)-hydroxy-L-ornithine, D-ala – D-alanine, abu- α -amino-isobutyric acid, aad- α -amino adipic acid, pip – pipercolic acid, hiv (hiv-d) - 2-hydroxy isovaleric acid (D form), bmt - (4R)-4-[(E)-2-butyl]-4-methyl-L-Thr, aoda- (S)-2-amino-8-oxodecanoic acid.

Our results for A-domain (full-length sequence) phylogenetic analysis are in line with similar studies done for mycotoxins (Gallo *et al.*, 2013). A-domains cluster together with other domains from the

same NRPS mega synthetase irrespective of their substrate specificity. A-domains belonging to PKS-NRPS or NRPS-like enzymes are clustered together according to their NRPS mega synthetase type or nature of biosynthesized chemical products. Because of this, full-length A-domain sequences would be of little help in decoding their substrate specificity.

3.2 Cluster analysis of NRPS code residues

Usually, full-length sequences are used to study phylogenetic relationships among A-domains. However, amino acid residues that are within a small distance ($<4\text{\AA}$) around substrate atoms (as determined by substrate-bound 3D enzyme structural complexes) determine specificity. For A-domains, these residues are defined as “NRPS code”. To study the similarity or differences in substrate specificity between bacteria and fungi, clustering of NRPS code residues was done (Fig 8).

3.2.1 Similarity/differences between fungal and bacterial NRPS codes

Cluster analysis of fungal and bacterial NRPS code residue shows differences for few substrates. This suggests that there could be an independent evolution of fungal and bacterial substrate specificity to bind the same substrate. There are some substrates for which there exist only a few characterized fungal A-domains and in turn fewer NRPS codes. Those A-domains and substrates have not been included in the previously developed tools for substrate specificity predictions and are underrepresented. Fungal A-domains are found to activate some special substrates that are rarely or not activated by bacteria e.g. phenyl-pyruvic acid, fumaric acid, 5-methyl-orsellinic acid. As these substrates are not well characterized in other A-domains, specificity predictions have been so far difficult.

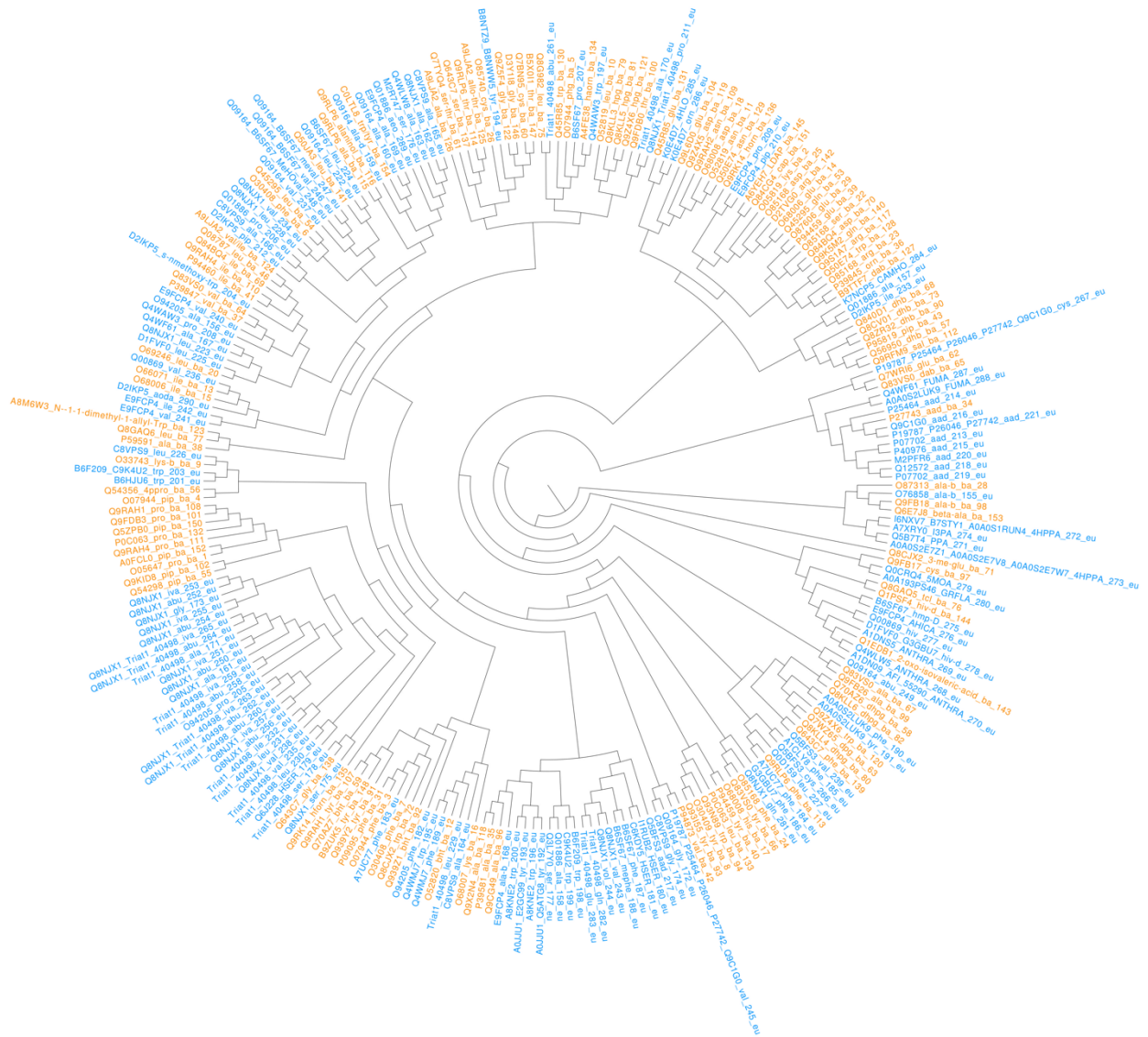


Figure 8. Bacterial and fungal NRPS code residue cluster analysis. Fungal (blue) and bacterial (orange) NRPS code residues were clustered after encoding them into physicochemical properties. Distinct clusters are obtained depending upon the type of substrate that binds. UniProt identifiers, substrate name, unique identifier and A-domain origin (ba- Bacterial and eu- Eukaryotic) are used to label the leaves of the tree.

Table 8. Results for cluster analysis of NRPS code residues. NRPS code residue similarity or differences between fungi and bacteria binding the same substrate are listed here.

Substrates	Clusters for bacterial and fungal NRPS codes
pro, pip	bacterial codes cluster together in a single clade, fungal codes do not cluster together and not with bacteria
gly, ser, homoserine	do not cluster together between bacteria and fungi
abu, iva	found in fungi, clustered together for these two substrates gly, pro, ala binding codes are interspersed between them
ala, leu, ile, val	codes are similar in fungi and bacteria
hiv, tcl, hmp-d, α -hydroxy-isocaproic acid	fungal and bacterial codes are in close proximity
phe	few codes are similar in bacteria and fungi
trp	well clustered in fungi but in bacteria, dispersed between different clades
aad	for ACVS synthetases, similar between bacteria and fungi
indole-3-pyruvic acid, 4-hydroxy-phenyl pyruvic-acid, phenyl-pyruvic acid	for these unusual substrates, clustered together in fungi
5-methyl-orsellinic acid, grifolic acid	cluster together in fungi

abu : α -amino-isobutyric acid, iva : isovaline, tcl : (4*S*)-5,5,5-trichloro-leucine, aad : α -amino adipic acid, pip : pipercolic acid, hiv : 2-hydroxy isovaleric acid.

3.3 Phylogenetic relationships among fungal C-domains

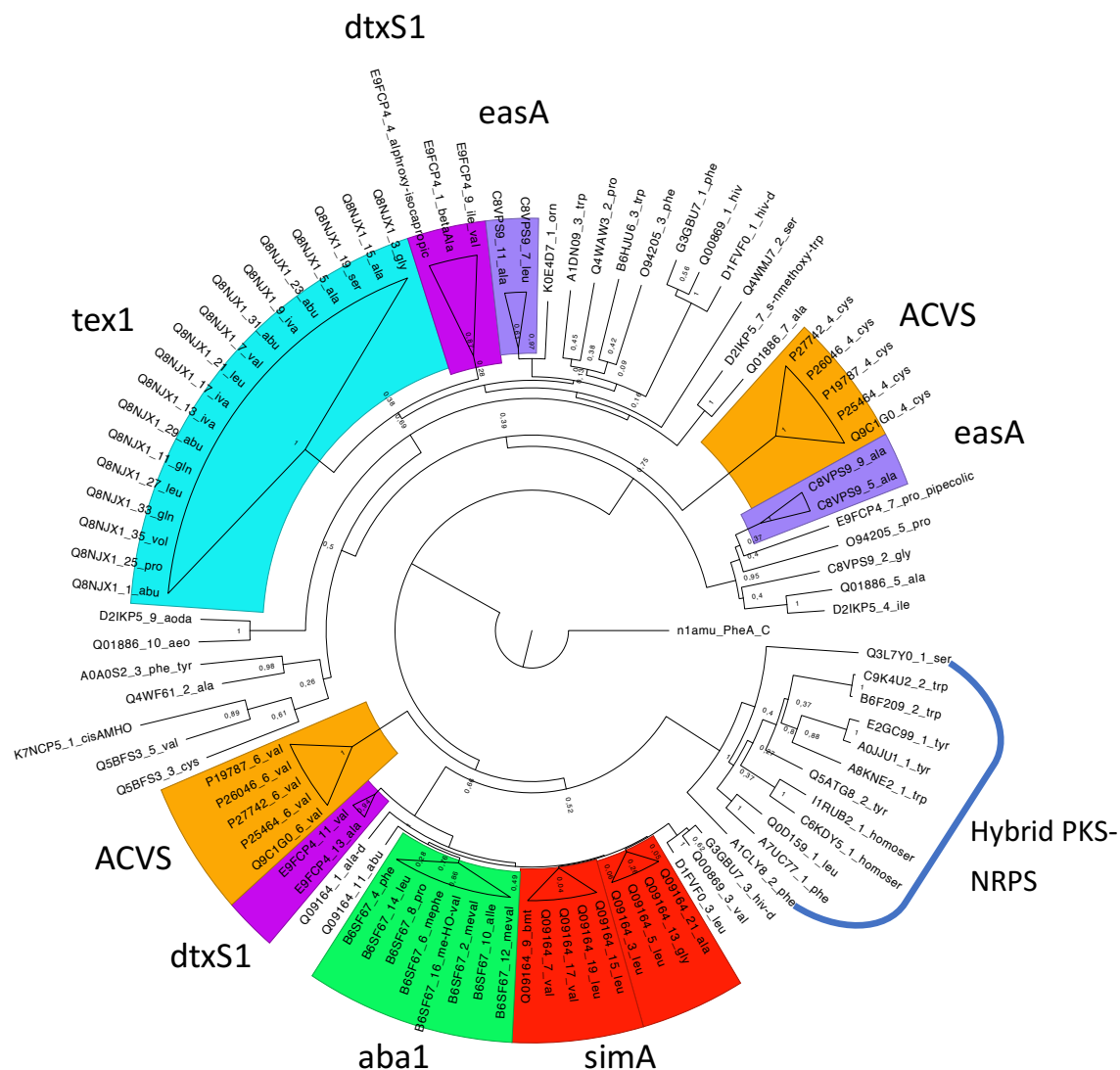


Figure 9. Fungal C-domain phylogenetic analysis. Leaves of the phylogenetic tree are labeled with substrate specificity information for downstream A-domain and UniProt identifier and number represents domain number obtained through Pfam domain architecture (Only A and C-domains were numbered in the entire NRPS sequence).

As opposed to A-domains, ACVS synthetase C-domains that are involved in condensation of val and cys do not cluster together. All A-domains of ACVS synthetases are clustered together in a single clade, while substrate specificity-wise clustering is maintained in the case of C-domains.

Hybrid PKS-NRPS synthetases C-domains are clustered together in a single clade and substrate specificity-wise clustering is observed.

All C-domains of NRPS (Peptaibols - Q8NIX1 and Aureobasidin A1 biosynthetic complex - B6SF67) are well clustered in a single clade. C-domain (K0E4D7) binding ornithine, clusters with a clade from C8VPS9 C-domains binding ala and leu. Cyclosporine synthetase (Cyclosporine synthetase - Q09164) C-domains show clustering in three smaller clades. C-domains from NRPS (B6HJU6 and O94205 – fungal ergot alkaloid) binding aromatic substrates trp and phe respectively, cluster together. C-domains (which may or may not bind to the same substrates) from the same NRPS mega synthetase do not cluster together, but clustering is observed for C-domains condensing similar substrates. This points to further investigation on finding key residues that contribute towards substrate specificity.

3.4 Clustering of NRP monomers and A-domain substrates

To find out structural similarities between A-domain substrates and NRP monomers, this clustering exercise was done. The dendrogram shown in Fig 10 could be used as reference when testing a set of possible substrates for enzymatic assays. Molecules having similar functional groups tend to cluster together in a single clade. It would also be expected that structurally or chemically similar substrates could bind the same A-domain with variable binding strengths. Accessory enzymes in NRPS mega synthetases modify incorporated monomers at any stage of peptide synthesis, hence monomers deduced from the final NRP structure could be different from substrates added by A-domains.

3.4.1 Modifications

Monomer diversity points towards the types of modifications that are previously found and could be expected in new NRPs. Accessory enzymes may bring about modifications (N or O methylation, formylation, acetylation, hydroxylation) to alter the substrate that is incorporated into the growing peptide chain. Structurally similar monomers that contain similar modifications are clustered together (Fig 10) e.g. N-formyl-isoleucine, N-formyl-leucine.

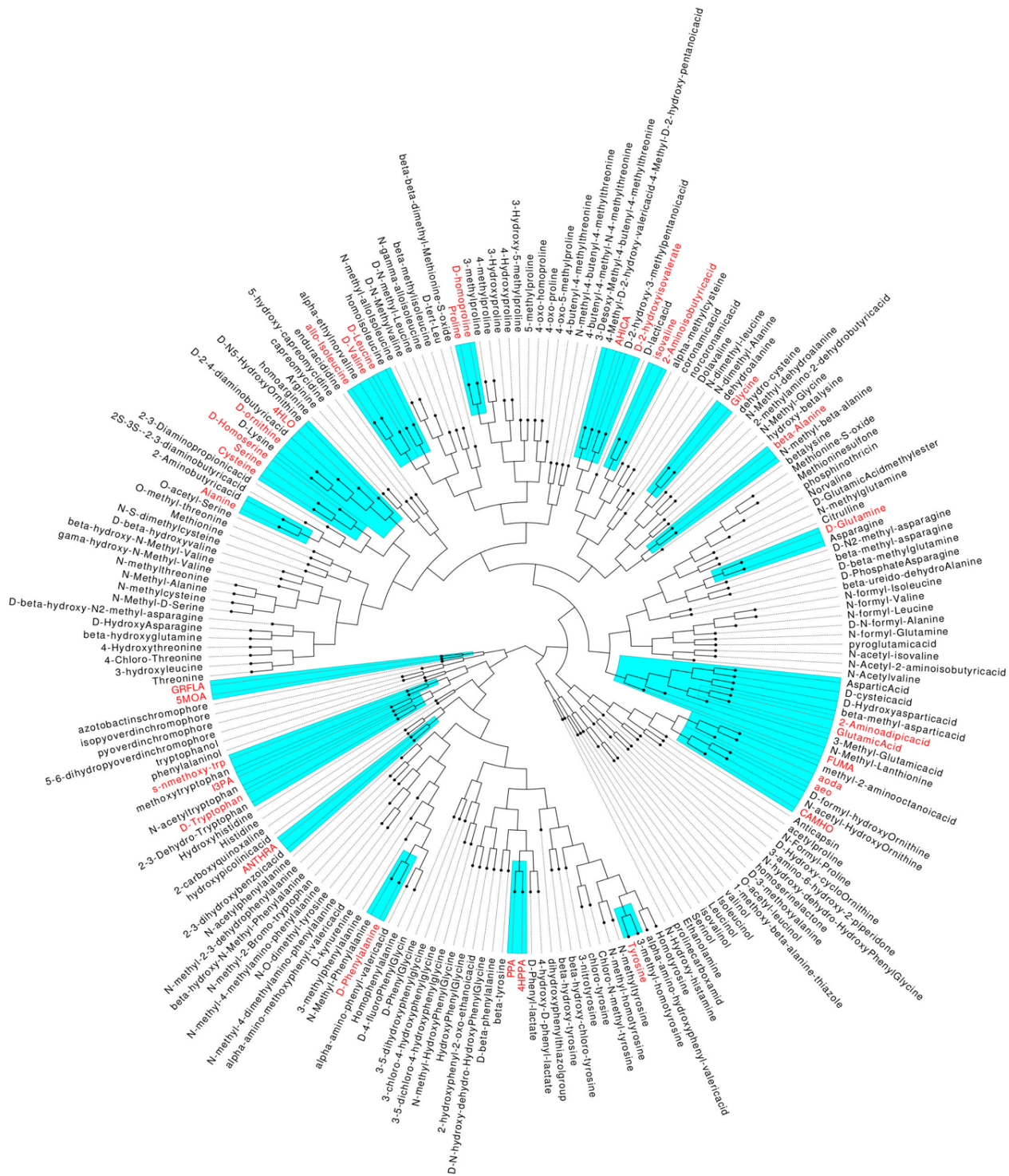


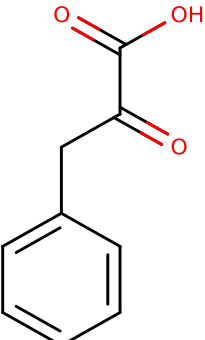
Figure 10. Structurally similarity of NRP monomers and A-domain substrates (whose NRPS codes are known). Substrates with at least one NRPS code identified through experimental characterization of fungal A-domains are colored in red and the closest node to the substrate is colored in cyan, to suggest a group of structurally similar substrates or monomers.

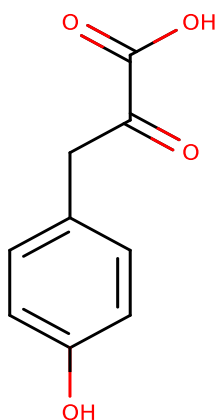
ANTHRA – anthranilic acid, 4HPPA- 4-hydroxy phenyl pyruvic acid, CAMHO - N(δ)-cis-anhydromevalonyl-N(δ)-hydroxy-L-ornithine, aeo - L-2-amino-8-oxo-9,10-decanoate, aoda - (*S*)-2-amino-8-oxodecanoic acid, FUMA - fumaric acid, 4HLO – 4-hydroxy-L-ornithine, AHICA - α -hydroxy-isocaproic acid, GRFLA - Grifolic acid, 5MOA – 5-methyl orsellinic acid.

3.4.2 A-domain substrates and NRPS code similarity

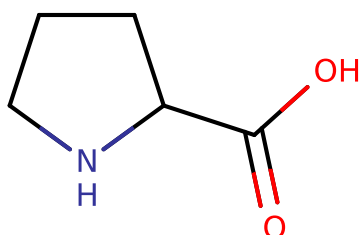
The rationale behind the A-domain substrate and NRP monomer cluster analysis was to see correlations between the structural similarity of substrates and NRPS code similarity. This correspondence was observed in few substrate pairs that are structurally very similar to each other (Table 9) e.g. proline and pipercolic acid, phenyl-pyruvic acid and 4-hydroxy-phenylpyruvic acid, 2-aminoisobutyric acid, and isovaline.

Table 9: Pair of structurally similar substrates and corresponding NRPS code residues. NRPS code residues that are similar between them are shown in boldface letters.

Structurally similar substrates	Substrate Structures	NRPS code residues
Phenylpyruvic acid		VATFIGGAG
4-hydroxy-phenylpyruvic acid		VAEFIGAAG

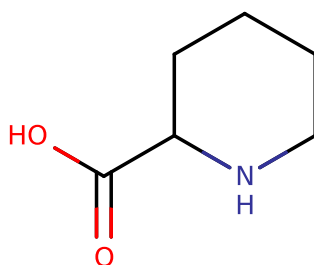


Proline



DI~~A~~VITVLI

Pipecolic acid



DVVALVLI

3.5 Substrate specificity predictions with NNassc

NRPS code residues were used as an input to NNassc and bit vectors with probability for the presence or absence of the substructures were predicted as an output. Tanimoto similarity coefficient values were calculated to match predicted bit vectors with probable training dataset substrate bit vectors. With decreasing values of coefficient, a ranked ordered list of probable substrates was obtained. The results obtained through NNassc were confronted with those with SANDPUMA.

3.5.1 Predictions - internal validation dataset

For 17 NRPS codes (73% of the internal validation dataset), correct substrates could be identified within the top five predictions (Table 10 and Appendix, Table 2, for a list of predictions for individual NRPS codes). Substrates with small size (aad, ala, pro, hiv-d, hmp-d, fumaric acid and abu) could be correctly predicted within top five predictions given by NNasc. However, some NRPS code residues binding substrates (phe, trp, gly and homoserine) the correct substrates could not be identified within the top five.

Table 10: NNasc and SANDPUMA substrate specificity prediction performance on the internal validation dataset (23 NRPS codes). SANDPUMA gives only one substrate prediction per NRPS code. Percentages of internal validation dataset substrates correctly identified within topmost ranked predictions for NNasc are listed. The number of NRPS codes out of the total dataset, the fraction corresponds is added into brackets.

SANDPUMA		NNasc	
Top 1	Top 1	Top 3	Top 5
30% (7)	56% (13)	65% (15)	73% (17)

3.5.2 Predictions - external validation dataset

Predictions for the external validation dataset (10 NRPS codes) are shown in Table 11. The top three predictions are given along with Tanimoto similarity values between the predicted fingerprint and closest substrate bit vector from the training dataset. The rank-ordered list of prediction is shown with decreasing values of Tanimoto similarity values. Our tool makes correct substrate predictions within the top three for all cases except for phenylalanine (NRPS code, DAYAVGGICK) and 4-hydroxyproline. In the case of 4-hydroxyproline, in the 2nd and 3rd place structurally similar substrates, proline and pipercolic acid respectively are found. SANDPUMA classifies only phenylalanine substrate correctly in one instance (NRPS code, DAYTSGGICK).

Table 11: External validation dataset predictions. NRPS codes are listed along with actual substrates and predictions from SANDPUMA and NNassc. For NNassc, the top three predictions are provided with Tanimoto similarity coefficient values (in brackets) between predicted and training dataset substrate fingerprints. When predictions match with the actual substrates, they are marked with boldface letters. pHMM – profile hidden markov model, ASM – active site motif.

NRPS code residues	NNassc predictions	SANDPUMA	Actual substrates
DVHHVTEIN	1. Ornithine (0.88) 2. Proline (0.74) 3. Pipecolic acid (0.7)	Alanine pHMM: abu-iva	4-hydroxyproline
DVHHVSGIN	1. Proline (1.0) 2. Pipecolic acid (0.95) 3. Ornithine (0.67)	Alanine pHMM: abu-iva	Proline
DVHQVSAIN	1. Proline (1.0) 2. Pipecolic acid (0.95) 3. Ornithine (0.67)	Alanine pHMM: abu-iva	Proline
DVHQVSGIN	1. Proline (1.0) 2. Pipecolic acid (0.95) 3. Ornithine (0.67)	Alanine pHMM: abu-iva	Proline
DIHQVSGIN	1. Ornithine (0.84) 2. Proline (0.8) 3. Pipecolic acid (0.76)	Alanine pHMM: abu-iva	Proline
GVIFIAAGI	1. Anthranilic acid (0.83) 2. Glycine (0.3) 3. Alanine (0.29)	Alanine pHMM: gln	Anthranilic acid
DVFFVVGVL	1. Alanine (0.92) 2. 2-amino-butyric acid (0.87) 3. Isoleucine (0.77)	Valine ASM: ala, SVM: val pHMM: aeo	Alanine
DAYAVGGIC	1. Alanine (0.85) 2. 2-amino-butyric acid (0.8) 3. Isoleucine (0.71)	Alanine pHMM: abu-iva	Phenylalanine
DLMLVGAVI	1. Alanine (0.86) 2. Isoleucine (0.82) 3. Leucine (0.71)	Alanine pHMM: gln	Leucine
DAYTSGGIC	1. Phenylalanine (0.95) 2. Tyrosine (0.83) 3. Alanine (0.55)	Phenylalanine SVM: phe, pHMM: gln	Tyrosine/Phenylalanine

NNasc differs from other A-domain substrate prediction tools by various aspects that are described below.

3.5.3 Structural insights into substrate specificity prediction

As there is not enough 3D structural information (PDB ID: 3ITE, only one solved eukaryotic A-domain crystal structure) for fungal A-domains, no structure-based prediction tool has been developed yet. A-domain specificity prediction by molecular docking was attempted by Lee *et al.* 2010, although they concluded that the accuracy of the 3D structural models imposes a limitation on deciphering correct substrates. Hence, concerted efforts towards crystalizing A-domains would help in modeling structures for homologous sequences through template-based modeling tools e.g. MODELLER (Eswar *et al.* 2006). 3D structures of A-domain-substrate complexes are required to map atomic-level interactions between substrates and binding site residues. For the working of NNasc, by the inclusion of structural properties of substrates and NRPS code amino acids in the form of Morgan fingerprints, attempts are made to fill gaps of structural information.

3.5.4 Prediction of substrate bit vectors

Previously developed A-domain substrate specificity prediction tools such as NRPSpredictor2, SANDPUMA suggest substrate class or group of substrates as outputs. Here, substrates are represented as multilabel vectors encoded into Morgan fingerprints; multiple labels are nothing but substructures of the molecule. By comparing the presence or absence of molecular substructures, one could find structurally similar molecules. Prediction of their molecular fingerprints of metabolites from their tandem mass spectra has been previously deployed by Dührkop *et al.*, 2015. In the future, larger databases of carboxy group-containing molecules could be collated and used to match probable substrates. This bit vector presentation for substrates is better than substrate classes, as these allow us to find novel substrates that share molecular substructures. Morgan fingerprint was chosen because they allow the generation of molecular substructures dynamically, rather than obtaining them from the predefined library of substructures.

NNassc accuracy could be further improved by considering the factors mentioned below.

3.5.5 Inclusion of more training data

Aromatic substrates (Phenylalanine and Tryptophan) have a larger size and could interact with most residues of NRPS code. Glycine and Homoserine have a smaller size, hence they can interact directly with only a few NRPS code residues. For these two substrate classes, NNassc could not infer correct substrates, but predictions could be improved by including more NRPS codes that represent different modes of binding.

The inclusion of more data for every substrate type would help improve the overall accuracy of NNassc. Machine learning algorithms depend on a huge amount of data to train their algorithms, the inclusion of more training examples always helps in building robust models.

3.5.6 Extraction of correct NRPS code residues

NNassc solely relies on nine residues of the NRPS code hence use of precise NRPS code is paramount to assigning correct substrates. When defining the NRPS code, Marahiel and coworkers used bacterial PheA structure, but for many fungal A-domains, homologous sequences (with known substrates) are used to extract NRPS codes. Given that A-domains share sequence identity in the range of 10-40%, finding their homologs through sequence alignments is uneasy. Hence, sequence alignment-free methods could be synergistically used with sequence-based methods e.g. profile HMM, etc. to identify remote homologs of A-domains (Guillermin *et al.* 2013) and subsequently extracting NRPS code residues precisely.

4. Conclusions

To assign a function for newly discovered A-domains, a phylogenetic analysis could be useful in certain scenarios. Here, phylogenetic analysis of A-domains was performed, which showed clustering by NRPS mega-synthetase or biosynthesized product type. Full-length A-domain phylogeny analysis did not give insights into substrate binding as clustering was observed for A-domains from the same NRPS megasynthetase. To understand relationships between A-domains and their substrate specificity, NRPS code residues were clustered. The clusters obtained show substrate specificity-wise clustering and suggests that, for few substrates, these residues are very different between fungi and bacteria. Along with A-domains, C-domains are also suspected to be involved in selecting substrates. Phylogeny analysis done here shows that for certain substrates, C-domains cluster together which share specificity.

As fungal and bacterial NRPS code do not share similarity for certain substrates, a novel approach for the prediction of fungal A-domain substrate specificity prediction is presented here. This involves predicting substrates using a combination of physicochemical properties and substructures of the NRPS code residues. Before this work on NNassc, there was no dedicated tool for fungal A-domain substrate specificity predictions. In spite of the scarcity of the data for fungal A-domain substrate specificity (only 136 NRPS codes and 41 substrates), correct predictions could be obtained for many NRPS codes. NNassc could be retrained when more fungal substrate specificity data becomes available in the future. As NNassc predicts substructures rather than substrates, novel substrates (not part of the training dataset) matching these properties could be suggested by NNassc. The performance of NNassc was evaluated on the internal validation dataset (23 NRPS codes) and external validation dataset (10 NRPS codes) to ensure that it gives reliable predictions for novel fungal A-domains. Though these validation datasets are small, they include quite diverse substrates. For the internal and external validation datasets, correct substrates could be predicted within the top five NNassc predictions for 17 and 8 NRPS codes respectively. Changes in A-domain substrate specificity, upon alterations in NRPS code residues, could also be predicted by NNassc when mutated residues are provided as an input. NNassc could be used to deduce substrate specificities for fungal A-domains and subsequently predict complete chemical structures of encoded peptides for newly sequenced fungal genomes.

5. References

1. AlQuraishi, M. AlphaFold at CASP13. *Bioinformatics* 1–4 (2019).
doi:10.1093/bioinformatics/btz422
2. Abadi, M, *et al.* Tensorflow: A system for large-scale machine learning. *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)* 265-283 (2016).
3. Bergendahl, V., Linne, U. & Marahiel, M. A. Mutational analysis of the C-domain in nonribosomal peptide synthesis. *Eur. J. Biochem.* **269**, 620–629 (2002).
4. Bergstra, J., Yamins, D. & Cox, D. D. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. *12th PYTHON Sci. CONF. (SCIPY 2013)* 13–20 (2013).
5. Blin, K. *et al.* AntiSMASH 4.0 - improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.* **45**, W36–W41 (2017).
6. Bloudoff, K., Alonzo, D. A. & Schmeing, T. M. Chemical Probes Allow Structural Insight into the Condensation Reaction of Nonribosomal Peptide Synthetases. *Cell Chem. Biol.* **23**, 331–339 (2016).
7. Boyle, N. M. O. *et al.* Open Babel: An open chemical toolbox. 1–14 (2011).
doi:10.1186/1758-2946-3-33
8. Bozhüyük, K. A. J. *et al.* De novo design and engineering of non-ribosomal peptide synthetases. *Nat. Chem.* **10**, 275–281 (2018).
9. Brakhage, A. A. Molecular regulation of beta-lactam biosynthesis in filamentous fungi. *Microbiol. Mol. Biol. Rev.* **62**, 547–85 (1998).
10. Chevrette, M. G., Aicheler, F., Kohlbacher, O., Currie, C. R. & Medema, M. H. SANDPUMA: Ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. *Bioinformatics* **33**, 3202–3210 (2017).
11. Caboche, S. *et al.* NORINE: A database of nonribosomal peptides. *Nucleic Acids Res.* **36**, 326–331 (2008).
12. Carpenter, K. A. & Huang, X. Machine Learning-based Virtual Screening and Its Applications to Alzheimer’s Drug Discovery: A Review. *Curr. Pharm. Des.* **24**, 3347–3358 (2018).

13. CHUN WEI YAP. PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* (2010). doi:10.1002/jcc
14. Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N. & Al Najada, H. Survey of review spam detection using machine learning techniques. *J. Big Data* **2**, (2015).
15. De Mattos-Shipley, K. M. J. *et al.* The cycloaspeptides: Uncovering a new model for methylated nonribosomal peptide biosynthesis. *Chem. Sci.* **9**, 4109–4117 (2018).
16. Donia, M. S. *et al.* A Systematic Analysis of Biosynthetic Gene Clusters in the Human Microbiome Reveals a Common Family of Antibiotics. *Cell* **158**, 1402–1414 (2014).
17. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl. Acad. Sci.* **112**, 12580–12585 (2015).
18. Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
19. Engel, T. Representation of chemical compounds. *Cheminformatics: A Textbook*, 15-168 (2003).
20. Epstein, S. C., Charkoudian, L. K. & Medema, M. H. A standardized workflow for submitting data to the Minimum Information about a Biosynthetic Gene cluster (MIBiG) repository: Prospects for research-based educational experiences. *Stand. Genomic Sci.* **13**, 1–13 (2018).
21. Eswar, N. *et al.* *Comparative Protein Structure Modeling Using Modeller*. *Curr Protoc Bioinformatics* (2006). doi:10.1002/0471250953.bi0506s15.
22. Flissi, A., Dufresne, Y., Michalik, J., Tonon, L., Janot, S., Noé, L., ... & Pupin, M. (2015). Norine, the knowledgebase dedicated to non-ribosomal peptides, is now open to crowdsourcing. *Nucleic acids research*, 44(D1), D1113-D1118.s
23. Frisvad, J. C. *et al.* Global analysis of biosynthetic gene clusters reveals vast potential of secondary metabolite production in *Penicillium* species. *Nat. Microbiol.* **2**, 17044 (2017).
24. Gallo, A., Ferrara, M. & Perrone, G. Phylogenetic study of polyketide synthases and nonribosomal peptide synthetases involved in the biosynthesis of mycotoxins. *Toxins (Basel)*. **5**, 717–742 (2013).

25. Gautam, G. & Yadav, D. Sentiment analysis of twitter data using machine learning approaches and semantic analysis. *2014 7th Int. Conf. Contemp. Comput. IC3 2014* 437–442 (2014). doi:10.1109/IC3.2014.6897213.
26. González-Medina, M. *et al.* Chemoinformatic expedition of the chemical space of fungal products. *Future Med. Chem.* **8**, 1399–1412 (2016).
27. Hyde, K. D. *et al.* The amazing potential of fungi: 50 ways we can exploit fungi industrially. *Fungal Divers.* (2019). doi:10.1007/s13225-019-00430-9
28. Heller, S. R., McNaught, A., Pletnev, I., Stein, S. & Tchekhovskoi, D. *InChI, the IUPAC International Chemical Identifier. Journal of Cheminformatics* **7**, (Journal of Cheminformatics, 2015).
29. Haynes, S. W., Ames, B. D., Gao, X., Tang, Y. & Walsh, C. T. Unraveling terminal C-domain-mediated condensation in fungal biosynthesis of imidazoindolone metabolites. *Biochemistry* **50**, 5668–5679 (2011).
30. Kalb, D., Lackner, G. & Hoffmeister, D. Functional and phylogenetic divergence of fungal adenylate-forming reductases. *Appl. Environ. Microbiol.* **80**, 6175–6183 (2014).
31. Kaljunen, H. *et al.* Structural Elucidation of the Bispecificity of A Domains as a Basis for Activating Non-natural Amino Acids. *Angew. Chemie - Int. Ed.* **54**, 8833–8836 (2015).
32. Keller, N. P. Fungal secondary metabolism: regulation, function and drug discovery. *Nat. Rev. Microbiol.* **17**, 167–180 (2019).
33. Chollet, F. & others, 2015. Keras. Available at: <https://github.com/fchollet/keras>.
34. Khaldi, N. *et al.* SMURF: Genomic mapping of fungal secondary metabolite clusters. *Fungal Genet. Biol.* **47**, 736–741 (2010).
35. Kudo, F., Miyanaga, A. & Eguchi, T. Structural basis of the nonribosomal codes for nonproteinogenic amino acid selective adenylation enzymes in the biosynthesis of natural products. *J. Ind. Microbiol. Biotechnol.* (2018). doi:10.1007/s10295-018-2084-7
36. Landrum, G. "RDKit: open-source cheminformatics software." (2016).
37. Lautru, S. & Challis, G. L. Substrate recognition by nonribosomal peptide synthetase multi-enzymes. *Microbiology* **150**, 1629–1636 (2004).

38. Larsson, J., Gottfries, J., Muresan, S. & Backlund, A. ChemGPS-NP: tuned for navigation in biologically relevant chemical space. *J. Nat. Prod.* **70**, 789–794 (2007).
39. Lee, T. V. *et al.* Structure of a eukaryotic nonribosomal peptide synthetase adenylation domain that activates a large hydroxamate amino acid in siderophore biosynthesis. *J. Biol. Chem.* **285**, 2415–2427 (2010).
40. Lee, T. V., Johnson, R. D., Arcus, V. L. & Lott, J. S. Prediction of the substrate for nonribosomal peptide synthetase (NRPS) adenylation domains by virtual screening. *Proteins Struct. Funct. Bioinforma.* **83**, 2052–2066 (2015).
41. Li, Y. F. *et al.* Comprehensive curation and analysis of fungal biosynthetic gene clusters of published natural products. *Fungal Genet. Biol.* **89**, 18–28 (2016).
42. Li, Z. R. *et al.* MODEL - Molecular descriptor lab: A web-based server for computing structural and physicochemical features of compounds. *Biotechnol. Bioeng.* **97**, 389–396 (2007).
43. Lipinski, A., Franc, I. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. **23**, (1997).
44. Macalino, S. J. Y., Gosu, V., Hong, S. & Choi, S. Role of computer-aided drug design in modern drug discovery. *Arch. Pharm. Res.* **38**, 1686–1701 (2015).
45. Martinez, J. L. General principles of antibiotic resistance in bacteria. *Drug Discov. Today Technol.* **11**, 33–39 (2014).
46. McNaught, A. The IUPAC International Chemical Identifier: InChI - A New Standard for Molecular Informatics. *Chem. Int.* 12–15 (2006).
47. Medema M. H. *et al.* Minimum Information about a Biosynthetic Gene cluster HHS Public Access Author manuscript. *Nat Chem Biol* **11**, 625–631 (2015)
48. Misiek, M., Braesel, J. & Hoffmeister, D. Characterisation of the ArmA adenylation domain implies a more diverse secondary metabolism in the genus *Armillaria*. *Fungal Biol.* **115**, 775–781 (2011).
49. Newman, D. J. & Cragg, G. M. Natural Products as Sources of New Drugs from 1981 to 2014. *J. Nat. Prod.* **79**, 629–661 (2016).
50. Patridge, E., Gareiss, P., Kinch, M. S. & Hoyer, D. An analysis of FDA-approved drugs: Natural products and their derivatives. *Drug Discov. Today* **21**, 204–207 (2016).

51. Rausch, C., Hoof, I., Weber, T., Wohlleben, W. & Huson, D. H. Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evol. Biol.* **7**, 1–15 (2007)
52. RDKit, Open-Source Cheminformatics. <http://www.rdkit.org>.
53. Reker, D. *et al.* Revealing the macromolecular targets of complex natural products. *Nat. Chem.* **6**, 1072–1078 (2014).
54. Reygaert, W. An overview of the antimicrobial resistance mechanisms of bacteria. *AIMS Microbiol.* **4**, 482–501 (2018).
55. Röttig, M. *et al.* NRSPredictor2 - A web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* **39**, 362–367 (2011).
56. Sandberg, M., Eriksson, L. & Sjö, M. New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. **2623**, 2481–2491 (1998).
57. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. 1–14 (2014).
58. Skinnider, M. A., Merwin, N. J., Johnston, C. W. & Magarvey, N. A. PRISM 3: Expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Res.* **45**, W49–W54 (2017).
59. Stachelhaus, T., Mootz, D. & Marahiel, A. The specificity-conferring code of adenylation nonribosomal peptide synthetases Torsten domains A Marahiel. *Elsevier Sci. Ltd* (1999).
60. Suzuki, R., Shimodaira, H., Hidetoshi, S. & Shimodaira, H. Package ‘ pvcust ’ Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling. 1–14 (2015).
61. Stratton, C. F., Newman, D. J., & Tan, D. S. (2015). Cheminformatic comparison of approved drugs from natural product versus synthetic origins. *Bioorganic & medicinal chemistry letters*, 25(21), 4802-4807.
62. Tang, G. W. & Altman, R. B. Knowledge-based Fragment Binding Prediction. *PLoS Comput. Biol.* **10**, (2014).
63. Töpfer, N., Fuchs, L. M. & Aharoni, A. The PhytoClust tool for metabolic gene clusters discovery in plant genomes. *Nucleic Acids Res.* **45**, 7049–7063 (2017).

64. Unson, M. D. & Faulkner D. J. Cyanobacterial symbiont biosynthesis of chlorinated metabolites from *Dysidea herbacea* (Porifera). *Experientia* **49.4**, 349–353 (1993).
65. Upson, R. H., Haugland, R. P., Malekzadeh, M. N. & Haugland, R. P. A spectrophotometric method to measure enzymatic activity in reactions that generate inorganic pyrophosphate. *Anal. Biochem.* **243**, 41–45 (1996).
66. Von Wintersdorff, C. J. H. *et al.* Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Front. Microbiol.* **7**, 1–10 (2016).
67. Wetzel, S. *et al.* Interactive exploration of chemical space with Scaffold Hunter. *Nature Chem. Biol.* **5**, 581–583 (2009).
68. Willighagen, E. L. *et al.* The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J. Cheminform.* **9**, 1–19 (2017).
69. Winn, M., Fyans, J. K., Zhuo, Y. & Micklefield, J. Recent advances in engineering nonribosomal peptide assembly lines. *Nat. Prod. Rep.* **33**, 317–347 (2016).
70. Wolf, T., Shelest, V., Nath, N. & Shelest, E. CASSIS and SMIPS: Promoter-based prediction of secondary metabolite gene clusters in eukaryotic genomes. *Bioinformatics* **32**, 1138–1143 (2016).
71. Zhang, T., Zhuo, Y., Jia, X., Liu, J., Gao, H., Song, F., ... & Zhang, L. (2013). Cloning and characterization of the gene cluster required for beauvericin biosynthesis in *Fusarium proliferatum*. *Science China Life Sciences*, *56*(7), 628-637.
72. Zobel, S. *et al.* Reprogramming the Biosynthesis of Cyclodepsipeptide Synthetases to Obtain New Enniatins and Beauvericins. *ChemBioChem* **17**, 283–287 (2016).

Appendix

Table 1. Fungal adenylation domain dataset, 116 NRPS code residues binding 41 different substrates. UniProt or similar identifier and 9 NRPS code residues along with experimentally characterized substrates are listed.

Identifier (UniProt)	9 NRPS code residues	Substrates
O76858	VDAVVSFGD	ala-b
O94205	DLFFCGGPL	ala
Q01886	DAGGCAMVA	ala
Q01886	DLLFGISVL	ala
Q09164	DLWFYIAVV	ala-d
Q09164	DVFIYAAIL	ala
Q8NJX1	DVGfVAGVL	ala
Q8NJX1	DIFVVAGVI	ala
Q4WLW8	DLFVLAGCI	ala
C8VPS9	DASQIGGIY	ala
C8VPS9	DLLVVAGIL	ala
C8VPS9	DIAILVAIL	ala
Q4WF61	DVYFTGGVL	ala
E9FCP4	DIFYAIATA	ala-b
E9FCP4	DVWIYA AVI	ala
#Triat1 40498	DILICALIC	ala
#Triat1 40498	DVGFLAGVF	ala
Q09164	DIQMfVAMQ	gly
Q8NJX1	DIGMVVGV L	gly
C8VPS9	DIQGV LAMQ	gly
Q8NJX1	DVGYLAAVY	ser
M2R747	DMWIAASIV	ser
Q3L7Y0	DLLMTWWIV	ser
#Triat1 40498	DVGYLMAVL	ser
Q6J228	DMTFVWGII	HSER
I1RUB2	DMTFVWGIN	HSER
C6KDY5	DMTFSWGIN	HSER
O94205	DLVGMAAVG	phe
A7UC77	DAYTMAAIC	phe
A7UC77	DAYTswAIC	phe
A1CLY8	DMSESWCFC	phe
G3GBU7	DGYCMAGAL	phe
B6SF67	DAWVLAGIQ	phe
Q4WMJ7	DAGTLGALM	phe
A0A0S2LUK9	DGYNAGSIC	phe
A0JJU1/Q5ATG8	DMVICGCAA	tyr
A0JJU1/E2GC99	DMVITWCAA	tyr
B8NTZ9/B8NWW5	DVFAFGAIF	tyr
A8KNE2	DMIICGCAA	trp
Q4WAW3	DVMFIGAVN	trp
B6F209	DMALAWSAC	trp
C9K4U2	DMALTWSAC	trp

A8KNE2	DMITWCAA	trp
B6HJU6	DIAMIGSMY	trp
#AFI_55290	DVMFVGEVA	trp
B6F209/C9K4U2	DMALCGSAC	trp
D2IKP5	DAFTLGCVF	s-nmethoxy-trp
O94205	DITLVAGLI	pro
Q01886	DIAVITVLI	pro
B6SF67	DVWVFSAIQ	pro
Q4WAW3	DVYFVGGIC	pro
E9FCP4	DLHEIGIIS	pro
Q8NXX1/#Triat1_40498	DVLFCLIC	pro
D2IKP5	DVVALVLI	pip
P07702	DPRHFVMRA	aad
P25464	EPRHIVEFV	aad
P40976	DPRHFVMRS	aad
Q9C1G0	EPRNVVEFV	aad
Q5BFS3	DPMMWMAIN	aad
Q12572	DPRHFVMIK	aad
P07702	DPRHFVMVK	aad
M2PFR6	DPRHFVMIP	aad
P19787/P26046/P27742	EPRNIVEFV	aad
Q09164	DAWLYGAVM	leu
Q8NXX1	DFLYFGGVV	leu
B6SF67	DAWMYGAVI	leu
D1FVF0	DGYIIGGVF	leu
C8VPS9	DIHFVGAIA	leu
Q0D159	DASLQWAIM	leu
#Triat1_40498	DFSYLGAVM	leu
#Triat1_40498	DAALVGVF	leu
#Triat1_40498	DMGWMGGVI	leu
D2IKP5	DAGGCSMVA	ile
Q8NXX1	DAALIGAVF	val
Q00869	DGWFIGIII	val
Q09164	DAWMFAAVL	val
Q8NXX1	DMGFLGGVC	val
Q5BFS3	DPLSTGAIG	val
E9FCP4	DAWFYGGTF	val
E9FCP4	DGLFIGIPV	ile
Q8NXX1	DAIIVGVT	val
P19787/P25464/P26046/P27742/Q9C1G0	DFESTAAY	val
Q09164/B6SF67	DAWMFAAIL	val
Q09164	DAWFHAVAY	abu
Q8NXX1	DLGFLAGLF	abu
Q8NXX1	DLGWLCGVF	abu
Q8NXX1	DCGWVVGVV	abu
Q8NXX1	DLGYLAGCF	abu
#Triat1_40498	DMGFIAGVV	abu
#Triat1_40498	DLGYVAGVF	abu
#Triat1_40498	DAFLLGIVA	abu
Q8NXX1/#Triat1_40498	DLGYLAGVF	abu
Q8NXX1/#Triat1_40498	DLGFLAGVF	abu
Q5BFS3	DVQHTITVV	cys
P19787/P25464/P26046/P27742/Q9C1G0	DHESDVGIT	cys
Q4WLW5	GVIIAAGI	ANTHRA
A1DNS5	GIILGAAGI	ANTHRA

A1DN09/#AFI_55290	GALFFAAGV	ANTHRA
Q5B7T4	VATFIGGAG	PPA
I6NXV7/B7STY1/A0A0S1RUN4	VAEFSGGAC	4HPPA
A0A0S2E7Z1/A0A0S2E7V8/A0A0S2E7W7	VAEFIGAAG	4HPPA
A7XRY0	VAHFTGAAC	I3PA
B6SF67	GALLVGITV	hmp-D
E9FCP4	GANLIGATV	AHICA
Q00869	GALHVVGSI	hiv
D1FVF0/G3GBU7	GALMVVGSI	hiv-d
Q0CRQ4	GFLTAGHAI	5MOA
A0A193PS46	GFVTGGFPL	GRFLA
Q8NJX1	DGGMVGGNY	gln
#Triat1_40498	DAAILVGVG	gln
K7NCP5	DVGGGGVIG	cisAMHO
K0E4D7	DVMELSSIT	orn
Q4WF61	SARGTVSQL	FUMA
A0A0S2LUK9	SARDVGSQL	FUMA
Q01886	DVLLCTGIM	aeo
D2IKP5	DGLTCGVII	aoda

cisAMHO : N(δ)-cis-anhydromevalonyl-N(δ)-hydroxy-L-ornithine, D-ala : D-alanine, abu : α -amino-isobutyric acid, aad : α -amino adipic acid, pip : pipercolic acid, hiv : 2-hydroxy isovaleric acid (hiv-d, D isomer), bmt : (4R)-4-[(E)-2-butyl]-4-methyl-L-Thr, aoda : (*S*)-2-amino-8-oxodecanoic acid, meval : N-methyl-valine, mephe : N-methyl phenylalanine, hmp-D : D-2-hydroxy-3-methylpentanoic acid, aeo : L-2-amino-8-oxo-9,10-decanoate, ANTHRA : anthranilic acid, 4HPPA : 4-hydroxy phenyl pyruvic acid, FUMA : fumaric acid, 4HLO : 4-hydroxy-L-ornithine, AHICA : α -hydroxy-isocaproic acid, GRFLA : Grifolic acid, 5MOA : 5-methyl orsellinic acid, HSER : homoserine, ala-b : beta-alanine.

Table 2. Internal validation dataset predictions. Substrate predictions by SANDPUMA and NNasc are listed. Predictions by individual algorithms from SANDPUMA are given along with the name of an algorithm (ASM – active site motif, SVM – support vector machine, pHMM – profile hidden markov model). Top 5 substrate matches for the NNasc are given along with Tanimoto coefficient. The number represents rank obtained by descending Tanimoto similarity values and rank for correct substrate is also given.

Nine NRPS code residues	Actual substrates	SANDPUMA predictions	NNasc (rank ordered substrate predictions)
EPRNVVEFV	2-aminoadipic acid	2-aminoadipic acid ASM: aad	1. 2-aminoadipic acid (1.0) 2. Ornithine (0.67) 3. 2-aminobutyric acid (0.65) 4. Glutamine (0.64) 5. Leucine (0.6)

		SVM: aad	
		pHMM: aad	
EPRNIVEFV	2-aminoadipic acid	2-aminoadipic acid	<ol style="list-style-type: none"> 1. 2-aminoadipic acid (1.0) Ornithine (0.67) 2-aminobutyric acid (0.65) Glutamine (0.64) Leucine (0.6)
		ASM: aad	
		SVM: aad	
		pHMM: aad	
DGGMVGGNY	Glutamine	Glutamine	<ol style="list-style-type: none"> 2-aminobutyric acid (0.88) Alanine (0.8) Leucine (0.78) 4. Glutamine (0.74) 2-amino-8-oxodecanoic acid (0.72)
		ASM: gln	
		SVM: gln	
		pHMM: gln	
DGYCMAGAL	Phenylalanine	Alanine	<ol style="list-style-type: none"> Alanine (0.92) 2-aminobutyric acid (0.87) Leucine (0.77) Valine (0.75) Isoleucine (0.67)
		pHMM: leu	
16. Phenylalanine (0.5)			
DGYNAGSIC	Phenylalanine	Alanine	<ol style="list-style-type: none"> Alanine (0.85) 2-aminobutyric acid (0.8) Isoleucine (0.71) Leucine (0.71) Ornithine (0.71)
		ASM: dpg	
		pHMM: trp	
19. Phenylalanine (0.52)			
DMIICGCAA	Tryptophan	Leucine	<ol style="list-style-type: none"> Phenylalanine (0.95) Tyrosine (0.83) Alanine (0.55) 4. Tryptophan (0.53) Phenylpyruvic acid (0.52)
		SVM: leu	
		pHMM: gly	
DVMFVGEVA	Tryptophan	Leucine	<ol style="list-style-type: none"> Alanine (0.79) Ornithine (0.77) 2-aminobutyric acid (0.75) 2-aminoadipic acid (0.72) Glutamine (0.72)
28. Tryptophan (0.34)			
DMALCGSAC	Tryptophan	Alanine	<ol style="list-style-type: none"> 1. Tryptophan (1.0) Indole-3-pyruvic acid (0.67) Phenylalanine (0.56) Tyrosine (0.48)
		pHMM: gly	

			5. S-n-methoxy-trp (0.47)
SARGTVSQL	Fumaric acid	Alanine pHMM: trp	1. Fumaric acid (0.91) 2. Glycine (0.57) 3. Alanine (0.53) 4. Cinnamic acid (0.53) 5. Caffeic acid (0.48)
DLFFCGGL	Alanine	Alanine/Glycine ASM: ala SVM: ala pHMM: gln	1. Alanine (1.0) 2. 2-amino-butyric acid (0.8) 3. Valine (0.8) 4. Isoleucine (0.71) 5. Leucine (0.71)
DLFVLAGCI	Alanine	Alanine pHMM: abu-iva	1. Alanine (1.0) 2. 2-amino-butyric acid (0.8) 3. Valine (0.8) 4. Isoleucine (0.71) 5. Leucine (0.71)
DVGFLAGVF	Alanine	Alanine	1. Alanine (1.0) 2. 2-amino-butyric acid (0.8) 3. Valine (0.8) 4. Isoleucine (0.71) 5. Leucine (0.71)
DIQGVLAMQ	Glycine	Alanine SVM: ala pHMM: gln	1. Alanine (0.85) 2. Ornithine (0.71) 3. 2-amino-butyric acid (0.69) 4. Valine (0.69) 5. 2-aminoadipic acid (0.67)
9. Glycine (0.64)			
DMTFVWGII	Homoserine	Alanine pHMM: gly	1. Alanine (0.93) 2. 2-amino-butyric acid (0.87) 3. Isoleucine (0.77) 4. Leucine (0.77) 5. Valine (0.75)
7. Homoserine (0.63)			
DVWVFSAIQ	Proline	Alanine SVM: ala pHMM: leu	1. Proline (0.9) 2. Pipecolic acid (0.85) 3. Ornithine (0.75) 4. Alanine (0.58) 5. 2-amino-butyric acid (0.57)
DVYFVGGIC	Proline	Alanine pHMM: abu-iva	1. Proline (0.94) 2. Pipecolic acid (0.9) 3. Trans-4-hydroxy-Proline (0.65) 4. Ornithine (0.62)

			5. Alanine (0.53)
DVVALVVLI	Pipecolic acid	Alanine pHMM: pro	<ol style="list-style-type: none"> 1. Pipecolic acid (0.95) 2. Proline (0.9) 3. Ornithine (0.6) 4. Trans-4-hydroxy-Proline (0.56) 5. Alanine (0.5)
DMGFLGGVC	Valine	Alanine ASM: val pHMM: abu-iva	<ol style="list-style-type: none"> 1. 2-amino-butyric acid (0.95) 2. Alanine (0.86) 3. Leucine (0.82) 4. Isoleucine (0.72) 5. Ornithine (0.72)
6. Valine (0.71)			
DAWFYGGTF	Valine	Valine ASM: ile, leu, val SVM: val pHMM: leu	<ol style="list-style-type: none"> 1. Alanine (0.92) 2. 2-amino-butyric acid (0.75) 3. Valine (0.75) 4. Isoleucine (0.67) 5. Leucine (0.67)
DGLFIGIPV	Isoleucine	Alanine ASM: ile, val pHMM: gln	<ol style="list-style-type: none"> 1. Alanine (0.86) 2. 2-amino-butyric acid (0.81) 3. Isoleucine (0.72) 4. Leucine (0.72) 5. Ornithine (0.72)
DLGYVAGVF	2-amino-butyric acid	2-amino-butyric acid/ isovaline	<ol style="list-style-type: none"> 1. 2-amino-butyric acid (0.93) 2. Alanine (0.86) 3. Leucine (0.72) 4. Valine (0.71) 5. Isoleucine (0.63)
GALLVGITV	Hmp-d	Alanine	<ol style="list-style-type: none"> 1. α-hydroxy-caproic acid (0.82) 2. D-2-hydroxyisovalerate (0.71) 3. D-2-hydroxy-3-methyl-pentanoic acid (0.63) 4. Valine (0.45) 5. Alanine (0.44)
GALMVVGS	Hiv-d	Alanine ASM: hyv pHMM: abu-iva	<ol style="list-style-type: none"> 1. D-2-hydroxyisovalerate (1.0) 2. D-2-hydroxy-3-methyl-pentanoic acid (0.78) 3. α-hydroxy-caproic acid (0.68) 4. Valine (0.43) 5. Alanine (0.42)

Acknowledgements

I would like to express my sincere gratitude to **Dirk Hoffmeister** for accepting me as a doctoral student and being a supportive mentor. His immense knowledge and expertise in fungal secondary metabolism helped a great deal while working on my research project. Also, his attention to detail and insightful comments strengthened my thesis.

I am very thankful to **Ekaterina Shelest** for supervising (tirelessly) throughout the doctoral project and giving invaluable guidance through these years. She gave me the freedom to implement my ideas independently but always directed me in the right direction when I was sidetracked. Her positive attitude, motivating, inquisitive, and caring nature were contagious, which in turn were beneficial for my professional and personal growth.

Additionally, my thanks go to **Reinhard Guthke** and **Gianni Panagioutou** for being very supportive during my time in their research groups in Jena.

I would also like to thank many colleagues whose constant support was paramount to completing my research work. **Francesco** (being always friendly and encouraging), **Cervin** (critical reading of the thesis and interesting discussions during the breaks), **Peter** (brainstorming sessions during the development of a classifier), **Thomas** (help with the poster and computer related stuff). All the current and past members of systems biology and bioinformatics & applied systems biology group members for delightful coffee and lunch breaks.

I cannot thank enough wonderful friends I made in Jena (**Amol, Alessandra, Tilottama, Kohulan, Shraddha, Prerna, Samir, and Ravindra**). My weekends were comforting and fun, all because of them. These pleasing weekends reenergized me to work with the same enthusiasm the next week.

Last but not least I am very much thankful to my Parents (**Shobha & Changdev**), my sister (**Komal**), and my cousin (**Vijay**) for always being there for me. This journey would be impossible without their love, constant encouragement and emotional support.

Curriculum Vitae

Sagar Gore

Theodor-Neubauer-Straße 46B,
04318 Leipzig, Germany.

Education

Doctoral student in Bioinformatics 2016 - 2019

Friedrich-Schiller-University, Jena, Germany.

Advisors: Dr. Dirk Hoffmeister and Dr. Ekaterina Shelest

Master of Science in Biotechnology (five years integrated) 2008 - 2013

Savitribai Phule Pune University, Pune, India.

Research experience

Doctoral student at *Friedrich-Schiller-University (FSU), Jena, Germany. (2016 - 2019)*

Project: Pattern recognition methods for prediction of chemical structures of fungal secondary metabolites (CRC 1127 ChemBioSys)

Project fellow at *Indian Institute of Science Education and Research, Pune, India. (2013 - 2015)*

Project: Structural descriptors of protein-protein and protein-ligand binding sites and knowledge-based design of new interfaces and ligands (Welcome Trust/DBT India Alliance).

Mentor: Dr. Mallur Madhusudhan

Project student at *Indian Institute of Science Education and Research, Pune, India. (2012 - 2013)*

Mechanism of lesion recognition by DNA repair enzyme, ALKBH2: A Quantum mechanical approach. *Mentor: Dr. Mrinalini Puranik.*

Fellowships to pursue Ph.D. in India

- Bioinformatics National Certification Exam – All India Rank - 3
- Biotechnology Eligibility Test by Department of Biotechnology, India.

Oral and poster presentations

- Invited short talk at Keystone Symposia J1 Natural Products and Synthetic Biology: Parts and Pathways, Olympic Valley, California, USA. (January 2018)
- Short talk at Jena School for Microbial Communication (JSMC) symposium, Jena, Germany. (October 2017)
- Poster presentation at IEEE International Conference on Computational Intelligence in Bioinformatics and Computational Biology, Manchester, UK. (August 2017) (received the best poster award)
- Poster presentation at MiCom 2017, 6th International Conference on Microbial Communication for Young Scientists, Jena, Germany. (March 2017)

Courses and Summer schools

- Molecular Modeling with Schrödinger-Suite Workshop
Leibniz Supercomputing Centre, Garching, Germany. (October 2018)
- Short Courses in Chemoinformatics: Databases and QSAR models,
Strasbourg University, Strasbourg, France. (May 2017)
- 3rd European Data Science Summer School
Saarland University, Saarbrücken, Germany. (September 2016)

List of publications

- **Sagar Gore**, Ekaterina Shelest. "Applying neural network model for the prediction of fungal adenylation domain substrate specificity". In preparation.
- Daniel Berry, Wade Mace, Katrin Grage, Frank Wesche, **Sagar Gore**, Christopher L. Schardl, Carolyn A. Young, Paul P. Dijkwel, Adrian Leuchtmann, Helge B. Bode, and Barry Scott. "Efficient non-enzymatic cyclisation and domain shuffling drive pyrrolopyrazine diversity from truncated variants of a fungal NRPS " *Proceedings of the National Academy of Sciences*, (accepted on 11-11-2019).
- Dörfer, Maximilian, Daniel Heine, Stefanie König, **Sagar Gore**, Oliver Werz, Christian Hertweck, Markus Gressler, and Dirk Hoffmeister. "Melleolides impact fungal translation via elongation factor 2." *Organic & biomolecular chemistry* 17, no. 19 (2019): 4906-4916.

Declaration/ Selbstständigkeitserklärung

Ich versichere, dass die hier vorgestellten Forschungsarbeiten nach bestem Wissen und Gewissen Originale sind, die, sofern nicht anders angegeben, aus meinen eigenen Untersuchungen hervorgegangen sind. Ich habe diese Dissertation selbst verfasst und keine Textabschnitte von Dritten oder von eigenen Dissertationen verwendet, ohne sie als solche zu kennzeichnen, und habe alle Hilfen, persönlichen Mitteilungen und Quellen innerhalb der Arbeit anerkannt. Ich habe keine unerlaubte fremde Hilfe in Anspruch genommen. Es wurde weder ganz noch teilweise für einen Abschluss an der Friedrich-Schiller-Universität oder einer anderen Universität eingereicht. Die Promotionsordnung der Fakultät der Biowissenschaften der Friedrich-Schiller-Universität ist mir bekannt.

I certify that the research work presented here is, to the best of my knowledge and belief, original that resulted from my own investigations, except as acknowledged. I have composed and written this dissertation and I have not used any sections of text from a third party or from dissertations of my own without identifying them as such, and has acknowledged all assistance, personal communication, and sources within the work. I did not use unauthorized outside help. It has not been submitted either in part or as a whole for any other degree at Friedrich-Schiller University or any other university. I have read the rules and regulations for submission of dissertation at the Faculty of biological sciences from Friedrich-Schiller University.

04.08.2020, Jena.

Sagar Changdev Gore