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

Sagar Changdev Gore, Ekaterina Shelest
Systems Biology/Bioinformatics
Leibniz Institute for Natural Product Research and Infection Biology- Hans Knöll Institute (HKI),
Jena, Germany.
Email: sagar.gore@leibniz-hki.de

In many bacteria, fungi and plants, secondary metabolites (SMs) are produced as a part of defense mechanism against co-localizing microorganisms or to strengthen symbiotic relationships with their hosts. Non Ribosomal Peptides (NRPs) and Polyketides are two major SM classes and their corresponding megaenzymes for their synthesis are Non-Ribosomal Peptide Synthetases (NRPS) and Polyketide Synthases (PKS). Genome mining efforts have suggested many SM biosynthesis gene clusters (groups of co-regulated and co-expressed genes) in fungi without any knowledge about produced SMs [1]. Biosynthetic genes are organized into various modules, that work like assembly line catalyzing different enzymatic reactions to build complete SM. NRPS assembly line has core domains such as Adenylation (A), Condensation (C) and Phosphopantetheine carrier protein (PCP or T) domain and terminal thioesterase domain (TE). There are some accessory domains such as methyltransferases and epimerases which further introduce modifications to a growing or complete NRP. NRPS adenylation (A) domains are involved in recognition of substrates and their activation by ATP dependent adenylation. A domains are known to recognize multitudes of substrates - 533 are known so far as deposited in Norine database[2], those include D and L configurations of 20 natural amino acids. C domains are involved in linking two substrates together with a peptide linkage. C domains have also been reported to show moderate specificity towards substrate that is activated by downstream A domain [3]. NRPs are built from substrates activated by A domains by incorporating one substrate at a time. An accurate prediction of substrates for A and C domains from sequences would be a first step in deciphering the complete SM structure.

Currently available sequence based tools for A domain substrate specificity prediction work well with bacterial sequences but for fungal sequences there is still a need for an improvement [4, 5]. Some of these tools use NRPS codes for substrate specificity predictions. NRPS code was defined from A domain phenyalanine bound structure PDB code 1amu, as 10 residues which are indispensable for substrate specific binding and catalysis. Our goal of this study is to improve fungal A domain substrate specificity prediction by making use of ligand and protein three dimensional structural data and eventually to predict fungal SM chemical structure.

To accomplish this we started with substrate structures and annotated NRPS code clustering analysis. 533 substrates obtained from Norine database were transformed into 461 non redundant Simplified Molecular Input Line Entry System (SMILES). This redundancy is due to the D or L forms of substrates. SMILES were then encoded into Morgan fingerprints [6], which take into account a circular neighborhood of atoms. These molecules were then assessed for their similarity using Tanimoto coefficient. Dendrogram consisting of 461 substrates was built, which shows distinct clusters for various substrate classes e.g hydrophobic or polar or hydrohphilc. Also substrates with similar atomic neighborhood are clustered together. Rationale behind this substrate clustering was to find structurally similar substrates for which NRPS codes could be deduced.

546 labeled A domain sequences with their corresponding NRPS codes were obtained from NRPSpredictor2 dataset and were used to build a dendrogram. 9 NRPS code residues were encoded by their physico chemical properties essential for binding and clustered with pvclust package [7] in R. Here, we observed that binding sites for same substrates (e.g trp, ser and pro) from bacterial and fungal enzymes do not cluster together suggesting an independent evolution of fungal A domain specificity. This emphasizes a need for the development of fungal specific tool for specificity prediction.

To build a classifier specifically for fungal sequences, inductive and transductive support vector machine (TSVM) approaches were used. Semi supervised learning with TSVM was used to incorporate large amount of unlabeled data. NRPSpredictor2 dataset (bacterial and fungal sequences) that was built in 2011 was composed of 576 labeled and 5096 unlabeled A domain sequences. We added 35 more labeled sequences to it. Though NRPSpredictor2 has fungal specific classifier but predictions are made at gross physico-chemical properties level. We have defined classes at more detailed level i.e group of similar substrate define one class. 86 labeled eukaryotic A domain sequences were divided into 9 substrate classes by considering substrate and NRPS code similarity, which was obtained from our earlier clustering exercise. We also used 783 unlabeled eukaryotic A domain sequences for TSVM analysis. NRPS code amino acids were encoded into aaindex [8] values with size, hydrophobicity and electronic properties which define binding site. Leave One Out (LOO) cross validation was performed and we obtained an overall accuracy of 80% for these 9 substrate classes. Our ongoing efforts are directed towards incorporating fragment/ fragment interaction information and more sequence features from both A and C domains to improve overall specificity prediction.

Keywords— *Non Ribosomal Peptide Synthetase (NRPS), Transductive Support Vector Machine (TSVM), Adenylation (A) domains, Condensation (C) domains.*

REFERENCES

[1] Inglis Diane O., et al. "Comprehensive annotation of secondary metabolite biosynthetic genes and gene clusters of Aspergillus nidulans, A. fumigatus, A. niger and A. oryzae." *BMC microbiology* 13.1 (2013): 1.

[2] Caboche Ségolène, et al. "NORINE: a database of nonribosomal peptides." *Nucleic acids research* 36.suppl 1 (2008): D326-D331.

[3] Rausch Christian, Ilka Hoof, Tilmann Weber, Wolfgang Wohlleben, and Daniel H. Huson "Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution." *BMC Evolutionary Biology* 7.1 (2007): 78.

[4] Röttig, Marc, et al. "NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity." *Nucleic acids research* (2011): gkr323.

[5] Baranašić, Damir, et al. "Predicting substrate specificity of adenylation domains of nonribosomal peptide synthetases and other protein properties by latent semantic indexing." *Journal of industrial microbiology & biotechnology* 41.2 (2014): 461-467.

6] Rogers, David, and Mathew Hahn. "Extended-connectivity fingerprints." *Journal of chemical information and modeling* 50, no. 5 (2010): 742-754.

[7] Suzuki, Ryota, and Hidetoshi Shimodaira. "Hierarchical clustering with P-values via multiscale bootstrap resampling." *R package* (2013).

[8] Kawashima, Shuichi, et al. "AAindex: amino acid index database, progress report 2008." *Nucleic acids research* 36.suppl 1 (2008): D202-D205.