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

Sagar Gore, Leibniz Institute for Natural Product Research and Infection Biology, Jena, Germany.

Non ribosomal peptide synthetases (NRPS)

Cyclosporin A (Immunosuppressant)
Norine (A. Flissi et al. 2016)

Goals:
I. To decipher rules governing fungal adenylation domain substrate specificity.
II. Prediction of complete chemical structure of secondary metabolites encoded by biosynthetic gene cluster (BGC) sequences.

Fig1. NRPS assembly line like mechanism with three modules each comprising multiple domains (e.g. A-T or C-A-T-E).

A: Adenylation T: Thiolation C: Condensation TE: Thioesterase M. Winn et al. 2015
Fungal A domain substrate specificity prediction

A domain substrates and NRPS code

534 (Norine database: 876 curated and 310 putative non ribosomal peptides)

D/L forms of proteogenic and nonproteogenic amino acids

NRPS code: 10 residues indispensable for binding

Inductive and transductive SVM

NRPS code: Wold encoding

9 substrate classes

86 (labeled) and 783 (unlabeled) eukaryotic sequences

<table>
<thead>
<tr>
<th>Tool</th>
<th>Algorithm</th>
<th>Dataset</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>SANDPUMA (2017)</td>
<td>Decision tree</td>
<td>928 SQ (90 Fungal) (Labeled) 104 SB</td>
<td>Accuracy 0.84</td>
</tr>
<tr>
<td>NRPS predictor</td>
<td>Support vector machines</td>
<td>576 SQ (Labeled) 5096 SQ (Unlabeled) 75 SB</td>
<td>Bacteria F 0.94 Fungi F 0.84</td>
</tr>
</tbody>
</table>

SQ= Sequences, SB= Substrates

Wold encoding

Hydrophobicity
Size
Electronic properties

DAGTLGALMK  phe

S. Wold et al. 1987

Fig2. Phenylalanine activating domain of gramicidin synthetas PheA (Brevibacillus brevis) PDB code 1AMU. 8Å residues in surface representation (tan color). 10 residues: NRPS code (cyan colored patch).