Finding Common Structural Motifs in Natural Products

Konoplev Denis

Alexey Gurevich, *Center for Algorithmic Biotechnology, SPbSU*

Hosein Mohimani, *UC San Diego*

Natural Products (NP)

A **natural product** is a chemical compound or substance produced by a living organism.

Why Study Natural Products?

NPs examples

- Antibiotics
- Antivirals agents
- Antitumor agents
- Immunosuppressors
- Toxins

Different Classes of NP

- Peptidic Natural Products (PNPs)
  - Non-Ribosomal Peptides (NRPs)
  - Ribosomally synthesized and Posttranslationally modified Peptides (RiPPs)
- Non-Peptidic Natural Products
  - Polyketides
  - Saccharides
  - Other

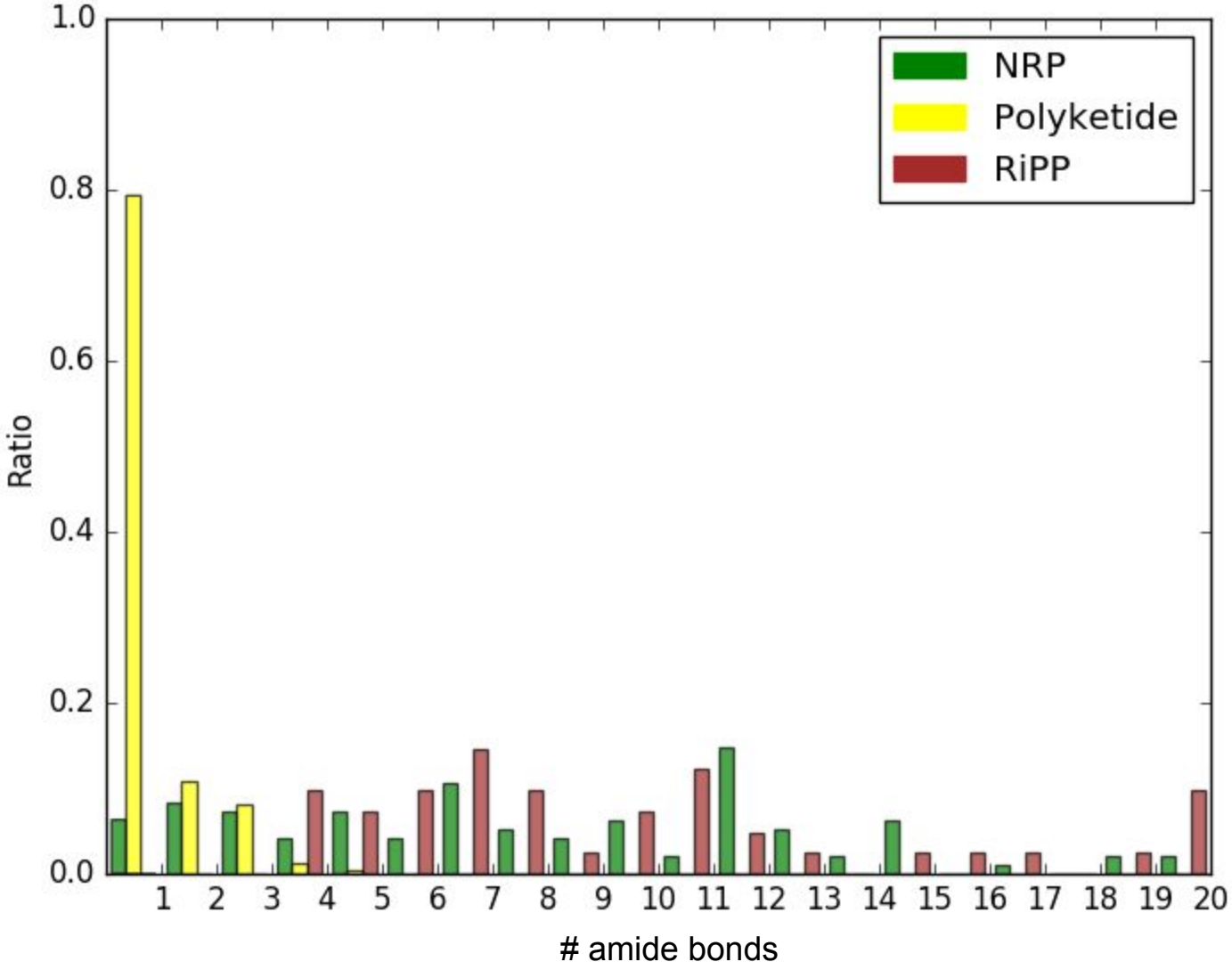
Goal and Objectives

Goal: develop a tool for Natural Products classification.

Objectives:

- Find common structural motifs and features for different classes of NP
- Develop a classification tool
- Validate the tool

Amide Bonds Distribution

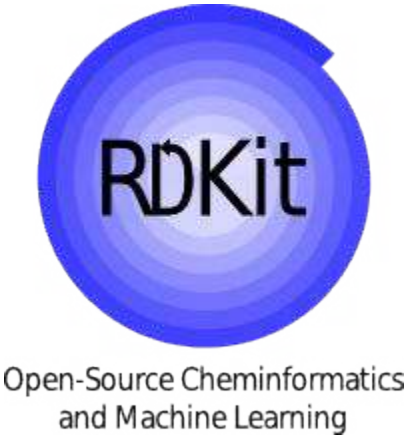


Finding Most Common Substructures

Idea: find the most common substructure in a class subset and count occurrences in other classes.

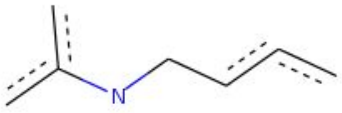
The default size of a subset - 10. The structure should occur at least in 90% of class structures.

Finding most common substructure and counting occurrences - RDKit Library

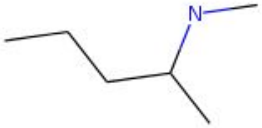


Finding Most Common Substructures

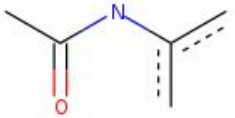
- We have found structural motifs for three most interesting classes of NPs: NRPs, RiPPs and Polyketides.
- These motifs are well represented (more than 90%) in one class and rarely occur (less than 50%) in other classes.



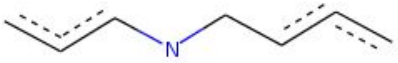
nrp



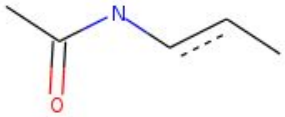
nrp



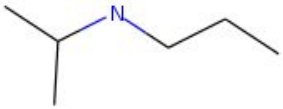
nrp



nrp

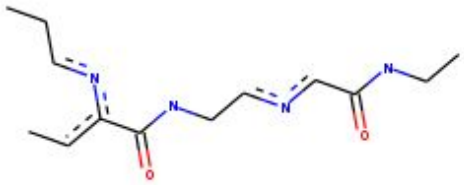


nrp

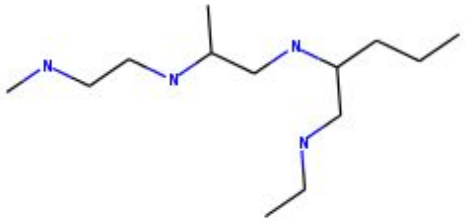


nrp

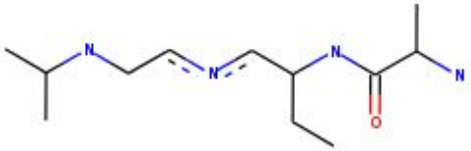
Most Common Substructures



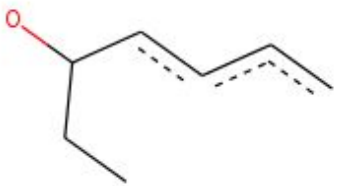
RiPP



RiPP

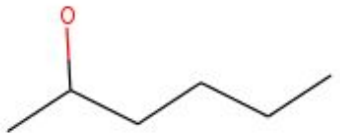


RiPP



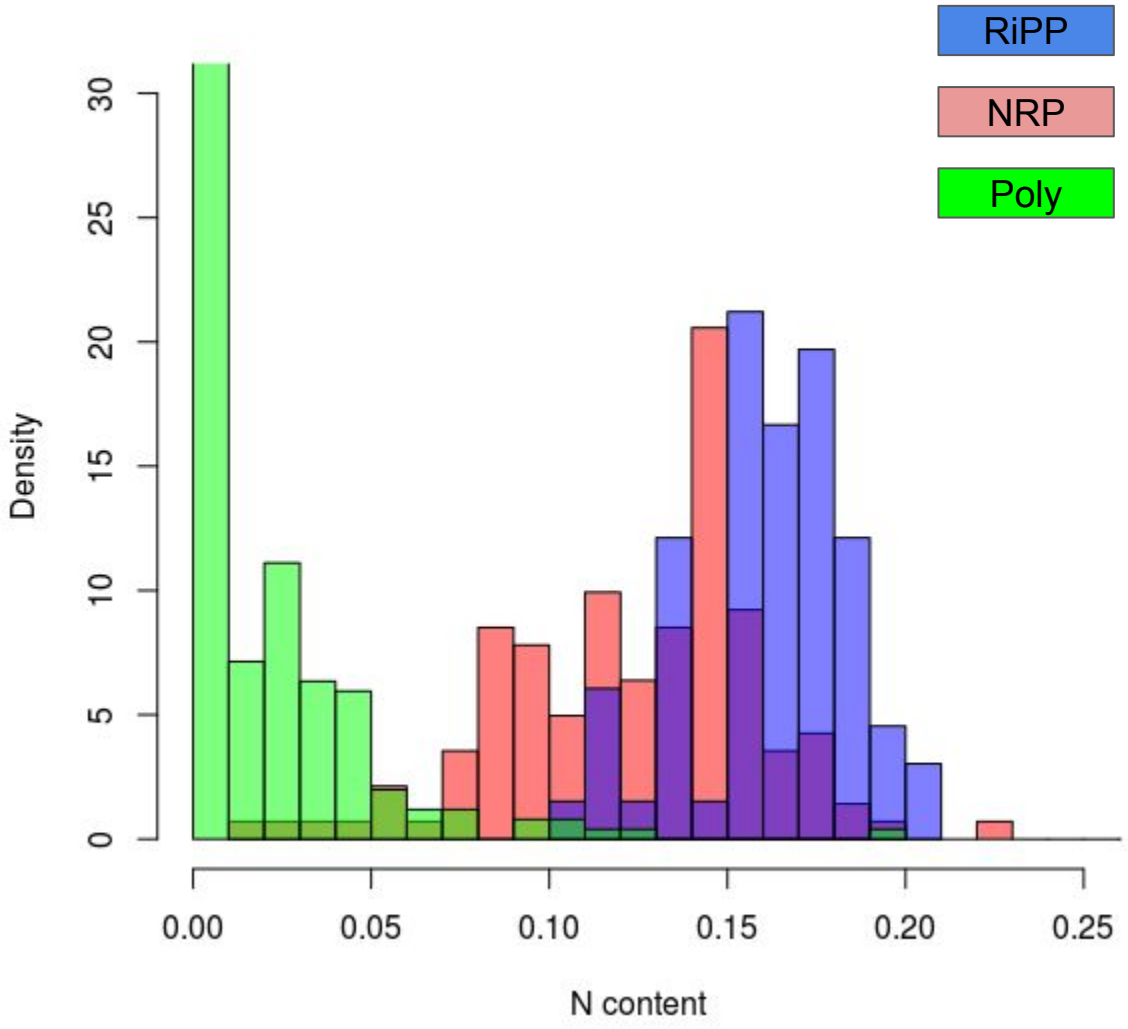
polyketide

polyketide



polyketide

Nitrogen content

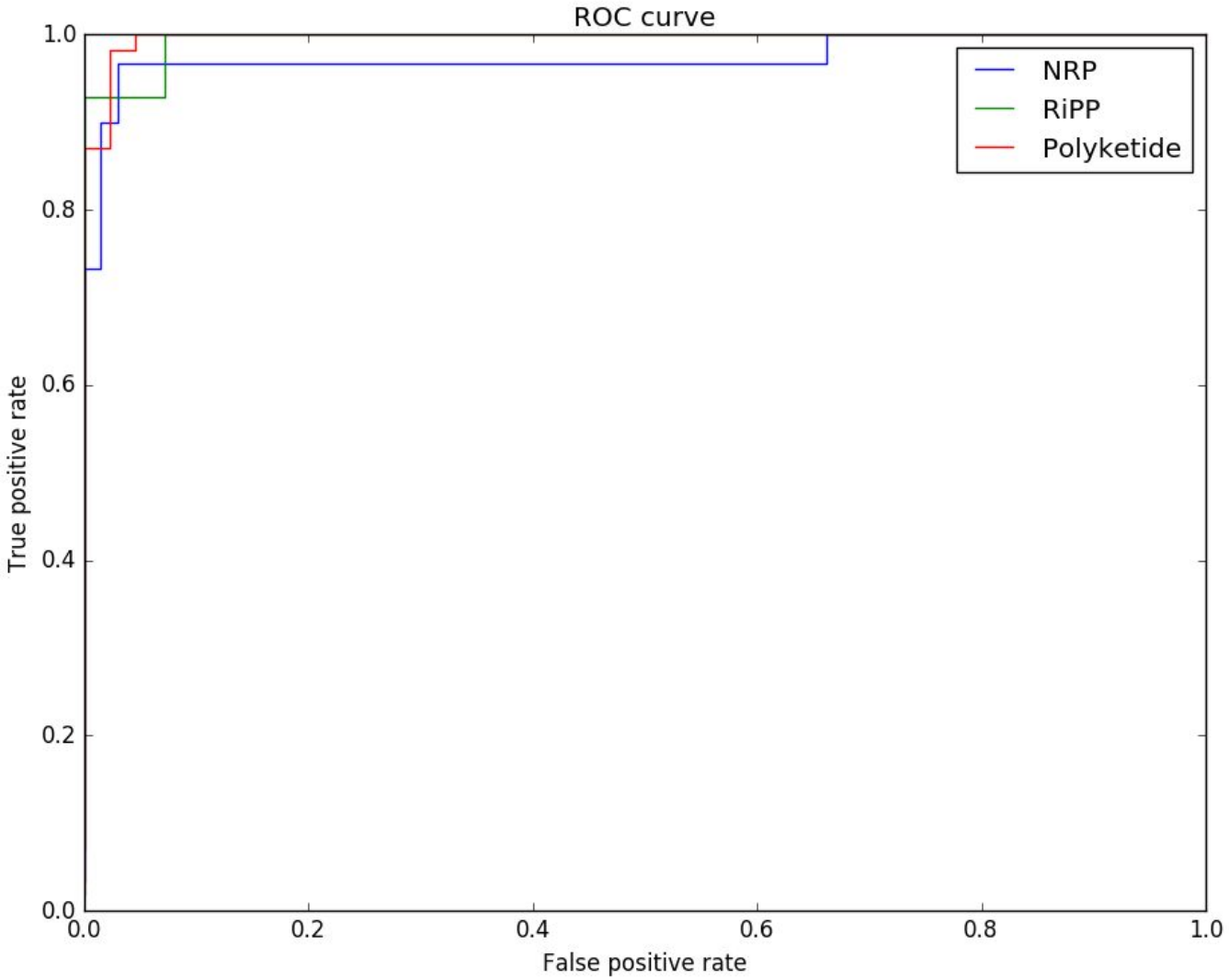


NP Classifier

- 15 features are structural motifs
- 3 are atom contents

Data is linearly separable therefore final model is Logistic regression.

ROC curve


ROC curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
NRP
RiPP
Polyketide

Precision Recall score

	Precision	Recall	F1-score	# structures
NRP	0.96	0.90	0.93	30
RiPP	0.93	0.93	0.93	14
Polyketide	0.96	1.00	0.98	54
avg/ total	0.96	0.96	0.96	98

Results

- Structural motifs for three classes of NPs were found
- Polyketides contain much less Nitrogens than other classes
- Tool for NPs classification was developed.

Problems

- Other classes
  - This classifier is not suitable for them
- Feature extraction is an unstable and expensive process

Thank You!