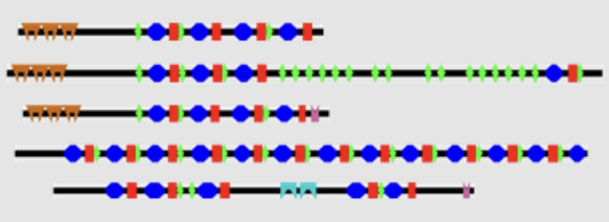


SPARCLE Curation

predicting protein architecture names

2024 NCBI Codeathon - Team Gwadz-Yang

Marc Gwadz (Team Leader), NCBI, NLM, NIH

Mingzhang Yang (Technical Lead), NCBI, NLM, NIH

Christopher Meyer (Writer), University of Chicago - Center for Translational Data Science

Franziska Ahrend, ORISE Fellow NIDDK, NIH

Yixiang Deng, Post-doc MIT, Harvard

Shaojun Xie, Advanced Biomedical and Computational Science, Frederick National Laboratory for Cancer Research

Goal

Input/processing

Model

Evaluation

Inference

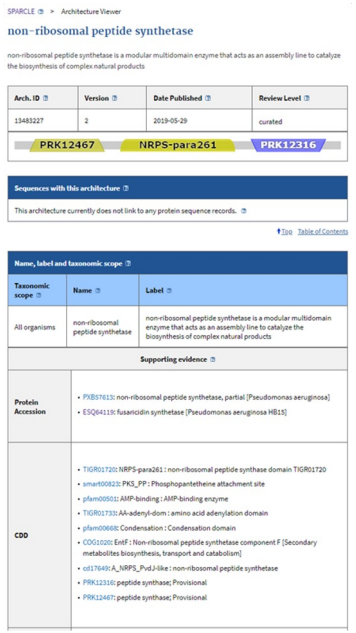
Future Work

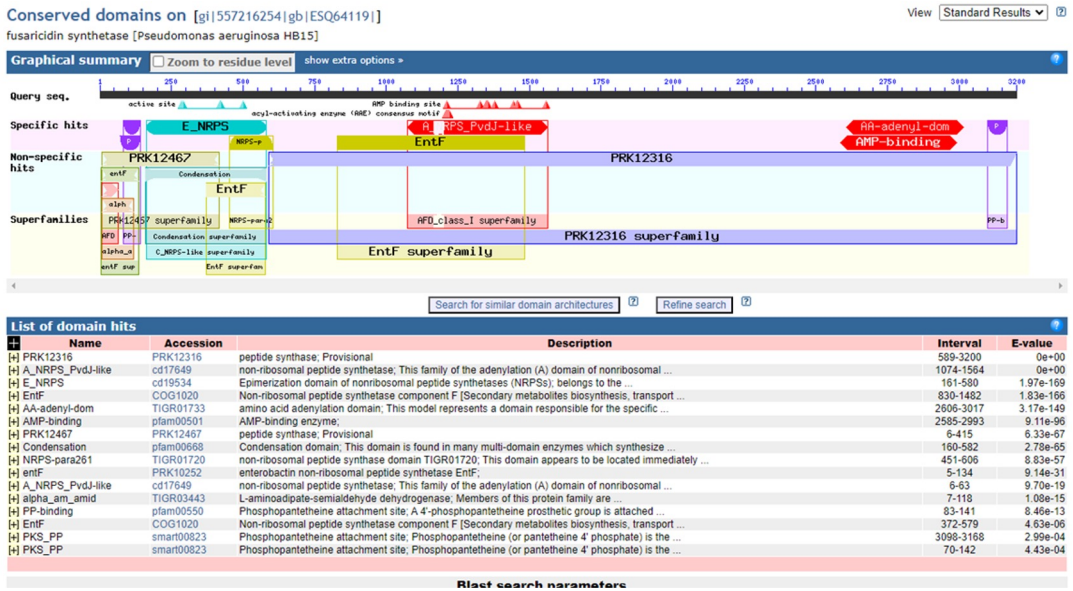
SPARCLE, the Subfamily Protein Architecture Labeling Engine, is a resource for the functional characterization and labeling of protein sequences that have been grouped by their characteristic conserved domain architecture.

A domain architecture is defined as the sequential order of conserved domains in a protein sequence.

Conserved domains can be clustered into superfamilies that generate over-lapping annotations on the same protein

Goal: Develop Machine Learning methods to use curated architectures to apply names (and other information) to un-curated architectures





Goal

Input/processing

Model

Evaluation

Inference

Future Work

Input Features: SuperFamilies, Specific Architectures, Title Strings
Output : Curated Names

Target output : Curated Names

Features to train Models: Superfamily and specific domain models that make up the architecture. Descriptive text associated with specific domains

Curated Name	LabelState	SpecificArch	superfamilyarch	TitleStrings
non-ribosomal peptide synthetase	curated	cl36129 TIGR01720 cl36106	NRPS-para261 PRK12316 PRK12467	PRK12467 | non-ribosomal peptide synthase domain TIGR01720 | PRK12316
NRPS-para261 and AFD_class_I domain-containing protein	namedByDomain	TIGR01720 cl36106	NRPS-para261 PRK12316	non-ribosomal peptide synthase domain TIGR01720 | PRK12316
NRPS-para261 and A_NRPS_Cytc1-like domain-containing protein	namedByDomain	cl36129 cl36106 TIGR01720 cl36106 cl36129 COG1020	NRPS-para261 PRK12316 PRK12467 EntF	PRK12467 | PRK12316 | non-ribosomal peptide synthase domain TIGR01720 | PRK12316 | PRK12467 | Non-ribosomal peptide synthetase component F [Secondary metabolites biosynthesis, transport and catabolism]
A_NRPS_Srf_like and A_NRPS domain-containing protein	namedByDomain	cl36129 COG3321 cl36129 cl11771	NRPS-para261 PRK12467 PksD	PRK12467 | Acyl transferase domain in polyketide synthase (PKS) enzymes [Secondary metabolites biosynthesis, transport and catabolism] | PRK12467 | NRPS-para261

Pre-processing :

- Curated Names were manually reduced to decrease # categories and help in testing.
- Common / uninformative words in TitleStrings (articles, prepositions, “protein” etc.) were eliminated prior to vectorization.
- Input/Output Data organized into One-Hot encoded matrices and vectors (SentencePiece)

Goal

Input/processing

Model

Evaluation

Inference

Future Work

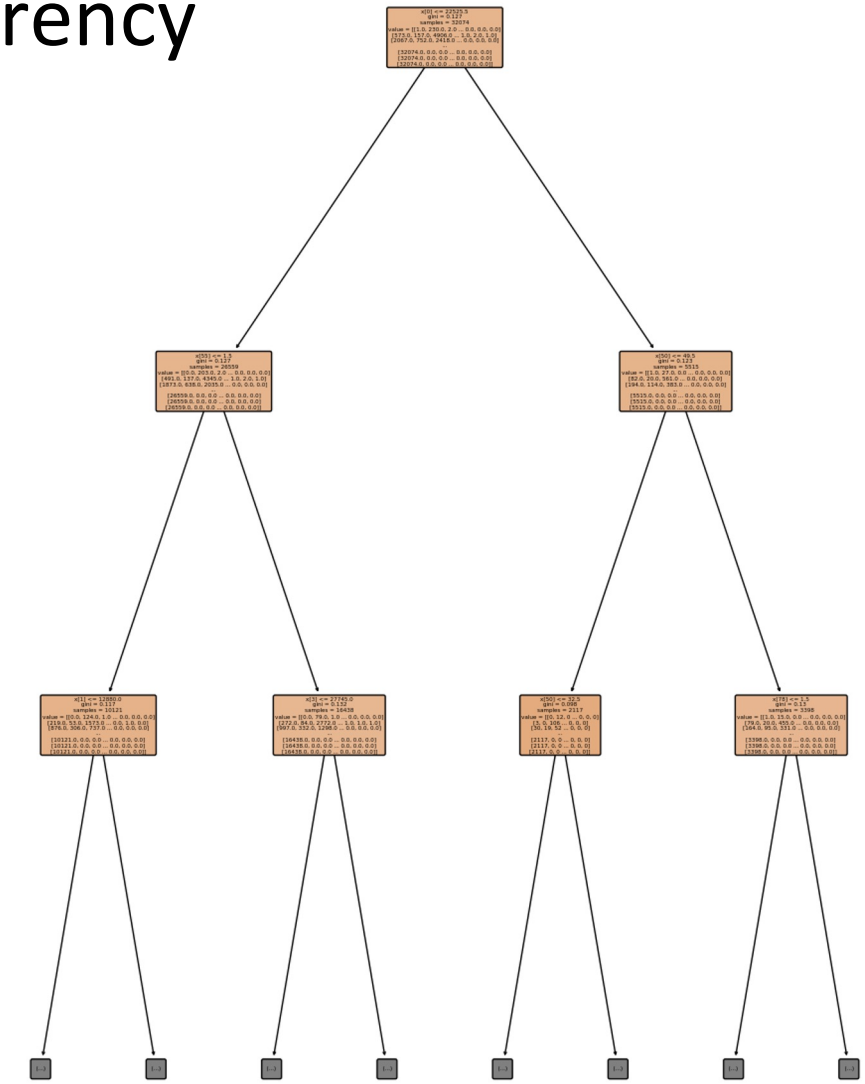
Preparation of the features and outputs

- Strategy A: One-Hot encoding for features
  - features size [48297] (Superfamilies, Specific CDs)
  - outputs as curated name strings
- Strategy B: Label encoding and Tokenization
  - feature size [450] (Superfamilies, Specific CDs, TitleStrings)

  - outputs as curated name strings or tokenized vector (size: 60)

Decision Tree

- Interpretability and Transparency
- Categorical Data Handling
- Reveals Feature Importance
- Models Non-Linear Relationships



- Goal
- Input/processing
- Model
- **Evaluation**
- Inference
- Future Work

Accuracy

Perfect match: >32%

Actual	Predicted	Equal
T-complex protein 1 subunit	T-complex protein 1 subunit	True
mannose-6-phosphate isomerase	mannose-6-phosphate isomerase	True
PUA domain-containing protein	PUA domain-containing protein	True
hybrid sensor histidine kinase/response regulator	hybrid sensor histidine kinase/response regulator	True
hybrid sensor histidine kinase/response regulator	hybrid sensor histidine kinase/response regulator	True
lysine decarboxylase	lysine decarboxylase	True
SANT/Myb-like DNA-binding domain-containing pr...	SANT/Myb-like DNA-binding domain-containing pr...	True
MFS transporter	MFS transporter	True
enoyl-CoA hydratase	enoyl-CoA hydratase	True
LCP family protein	LCP family protein	True
cell division protein	cell division protein	True
methyl-accepting chemotaxis protein	methyl-accepting chemotaxis protein	True
E3 ubiquitin-protein ligase	E3 ubiquitin-protein ligase	True
helix-hairpin-helix domain-containing protein	helix-hairpin-helix domain-containing protein	True
adenylate/guanylate cyclase domain-containing ...	adenylate/guanylate cyclase domain-containing ...	True
cadherin repeat domain-containing protein	cadherin repeat domain-containing protein	True
LacI family DNA-binding transcriptional regulator	LacI family DNA-binding transcriptional regulator	True
BTB/POZ domain and ankyrin repeat-containing p...	BTB/POZ domain and ankyrin repeat-containing p...	True
dTDP-glucose 4,6-dehydratase	dTDP-glucose 4,6-dehydratase	True
class I SAM-dependent methyltransferase	class I SAM-dependent methyltransferase	True
FRMD7 family protein	FRMD7 family protein	True
non-ribosomal peptide synthetase	non-ribosomal peptide synthetase	True
BTB/POZ domain-containing protein	BTB/POZ domain-containing protein	True
cation diffusion facilitator family transporter	cation diffusion facilitator family transporter	True
transposase	transposase	True

- Goal
- Input/processing
- Model
- **Evaluation**
- Inference
- Future Work

Accuracy

Perfect match: >32%

Similar match: 39%

Actual	Predicted
peroxisomal biogenesis factor	peroxisome biogenesis factor
methylmalonyl-CoA mutase subunit	methylmalonyl-CoA mutase
ubiquitin carboxyl-terminal hydrolase family p...	ubiquitin carboxyl-terminal hydrolase
C2HC-type zinc finger protein	C2H2-type zinc finger protein
AP-4 complex subunit	AP-3 complex subunit

Actual	Predicted
DUF4299 domain-containing protein	DUF4278 domain-containing protein
DUF4345 domain-containing protein	DUF4346 domain-containing protein
DUF3015 domain-containing protein	DUF2515 domain-containing protein
DUF5043 domain-containing protein	DUF5035 domain-containing protein

Actual	Predicted
ParB/RepB/Spo0J family partition protein	ROK family transcriptional regulator
flagellar brake protein	leucine--tRNA ligase
transposase	heat shock 70 family protein
GATOR complex protein	Myb family transcription factor
2-aminoethylphosphonate ABC transporter substr...	low molecular weight protein-tyrosine-phosphat...
PTS phosphocarrier protein NPr	PTS galactitol transporter subunit
DUF3083 family protein	YusG family protein
MspA family porin	class I SAM-dependent methyltransferase
CPBP family intramembrane glutamic endopeptidase	DUF4430 domain-containing protein
coproporphyrinogen III oxidase	NADH dehydrogenase subunit

Goal

Input/processing

Model

Evaluation

Inference

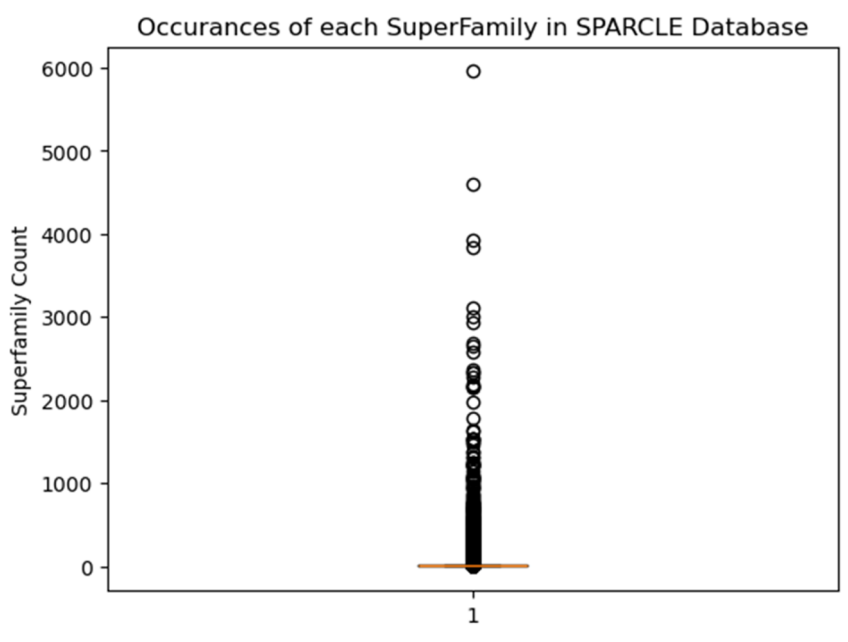
Future Work

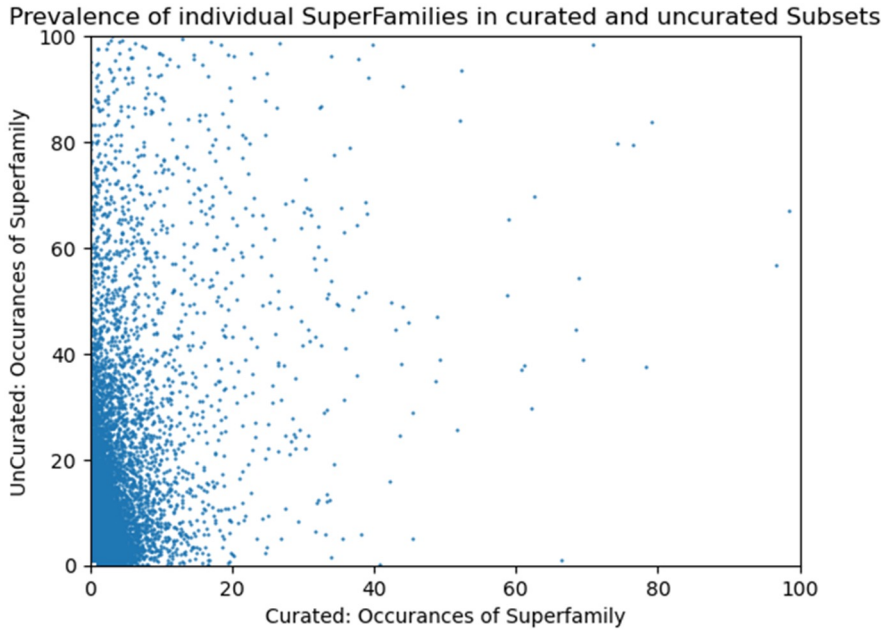
Challenges for training/testing :

- Sparse data (many architectures only have rare superfamilies)
- Large number of categories and some inconsistencies in Curated names

While most super superfamilies occur in very few Architectures, others can occur many times in both curated and uncurated Architectures.

Architectures with frequently occurring superfamilies can hopefully be leveraged to auto-name numerous related architectures.





Goal

Input/processing

Model

Evaluation

Inference

Future Work

- Standardized/generic target names
- Collect all the information / knowledge about every domain model to create a corpus.
  - With this corpus, we can generate word embedding for every single domain model.
  - Then we can create biologically meaningful vectors for the specificArch and superfamilyArch.
- Try other ML methods and employ better validation methods
- Parameter optimization: change “opt” variable to use different regressors.
  - We used “Ridge regression” (opt=ridge), a linear model with performance score $R^2 = 0.25$.
  - This is fastest, but would love to evaluate other methods.
- Evaluate choice and filtering of input features

  e.g. limiting text to informative strings
- Curator inspection of proposed names for uncurated data

#team-gwadz-yang

Team Leader
Marc Gwadz

Technical Lead
Mingzhang Yang

Writer
Christopher Meyer

Franziska Ahrend

Yixiang Deng

Shaojun Xie