Biosynthetic Gene Clusters classification

Lucie Cervenkova

CTU - FIT

cervel10@fit.cvut.cz

December 2022

Introduction

Natural products represent a reservoir of bioactive compounds with significant antimicrobial, anticancer and immunomodulatory effects. They are produced by bacterial and fungal organisms as their secondary metabolites, which are synthesized by sets of co-localized genes termed BGCs (Biosynthetic Gene Clusters). A rise of Next Generation Sequencing (NGS) methods led to the need of development of BGC prediction algorithms such as DeepBGC (Hannigan, et al., 2019), antiSMASH (Medema, et al., 2011) and ClusterFinder (Cimermanic, et al., 2014). A common postprocessing step in this prediction is a BGC product classification. The aim of this work is to explore deep learning strategies in the product classification with utilization of different embedding methods introduced in DeepBGC.

Input data

Training and validation dataset was derived from MIBiG database version 3.0 (Kautsar, et al., 2020), which records all confirmed BGCs. Data was downloaded as multiple GenBank files, which contain sequential information as well as important metadata (e.g. Pfam domains). These files were then converted in a desired tabular format utilizing `deepbgc prepare` command.

Preprocessing

DeepBGC improves its prediction step by taking position dependency of individual genomic elements into account. Their approach considers Pfam (protein family) domain membership of each gene cluster. In this work, two BGC representations were explored. First, a simple one-hot encoding representation, which is used in classification step of the latest DeepBGC version. Second, Pfam2Vec embedding was introduced. It was generated by the original word2vec method, where the training body consisted of 3 376 bacteria (documents) and 15 686 unique Pfam identifiers (words).

Methods

DeepBGC utilizes RandomForestClassifier as its current classification model. Before exploring deep learning approaches number of sklearn models, which support multi-label classification, were used (see results in Table 3). Afterwards, multiple layer perceptrons were built using Tensorflow/Keras libraries. I explored 8 different architectures (Table 1) with various number of hidden layers and number of neurons in them. Each neural network also contained dropout layer in between each hidden layer with set value of 0.2. To alleviate possible overfitting, early stopping was implemented with patience 10, which stops the training phase if the training loss in no longer decreasing after 10 epochs straight. Binary cross-entropy was used as loss metric, together with sigmoid activation function in output layer. These models were applied to input data in both formats – One-Hot Encoding, Pfam2Vec embedding. Evaluation was carried out using 3 times repeated 5-fold cross-validation with random splits (sklearn RepeatedKFold).

Model	Hidden Layers	Nodes per layer
NN1	1	100
NN2	2	100
NN3	3	100
NN4	4	100
NN5	3	50
NN6	2	50
NN7	1	50
NN8	1	10

Table 1: Deep learning models

Biosynthetic Gene Clusters classification

Lucie Cervenkova

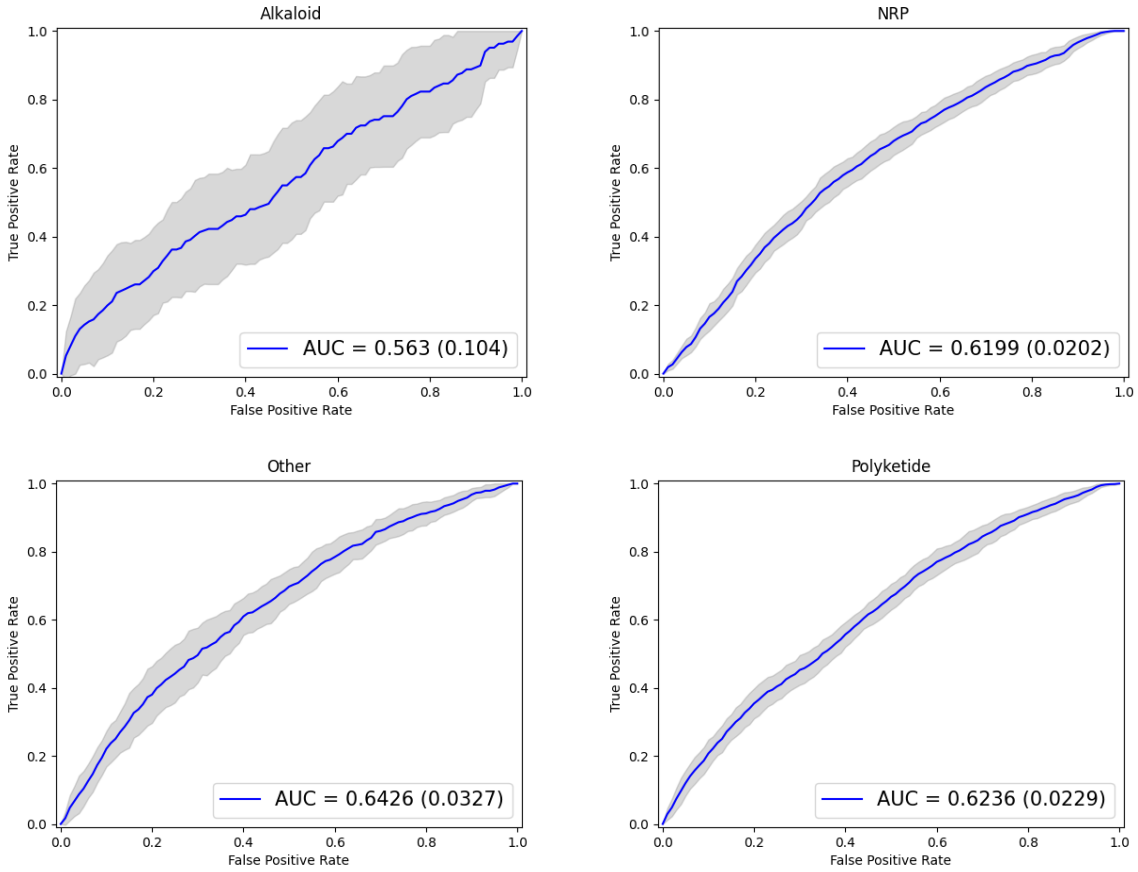
CTU - FIT

cervel10@fit.cvut.cz

December 2022

Results and discussion

Neural networks, unfortunately, did not beat existing RandomForestClassifier. While RandomForestClassifier perfoms well on some labels and worse on others (e.g. alkaloids), multiple layer perceptrons classify with similar score on all labels. Furthermore, I have encountered major issues with overfitting, especially with input in one-hot encoding format, which led me to conclude that multiple layer perceptron is in fact not suitable for this task. Among the attempted approaches, neural networks with 100 neurons in their hidden layers perform the best, when Pfam2Vec is used as an embedding, however their AUC significantly fluctuates in different cross-validation training subsets. In my opinion, high AUC values for sklearn's MLPClassifier are caused due to its inability to properly prevent overfitting. To sum up, multiple layer perceptron is not able to confidently classify BGC products. The only model that might be worth exploring for future DeepBGC updates would be ExtraTreesClassifier that outperformed existing RandomForestClassifier.



Biosynthetic Gene Clusters classification

Lucie Cervenkova

CTU - FIT

cervel10@fit.cvut.cz

December 2022

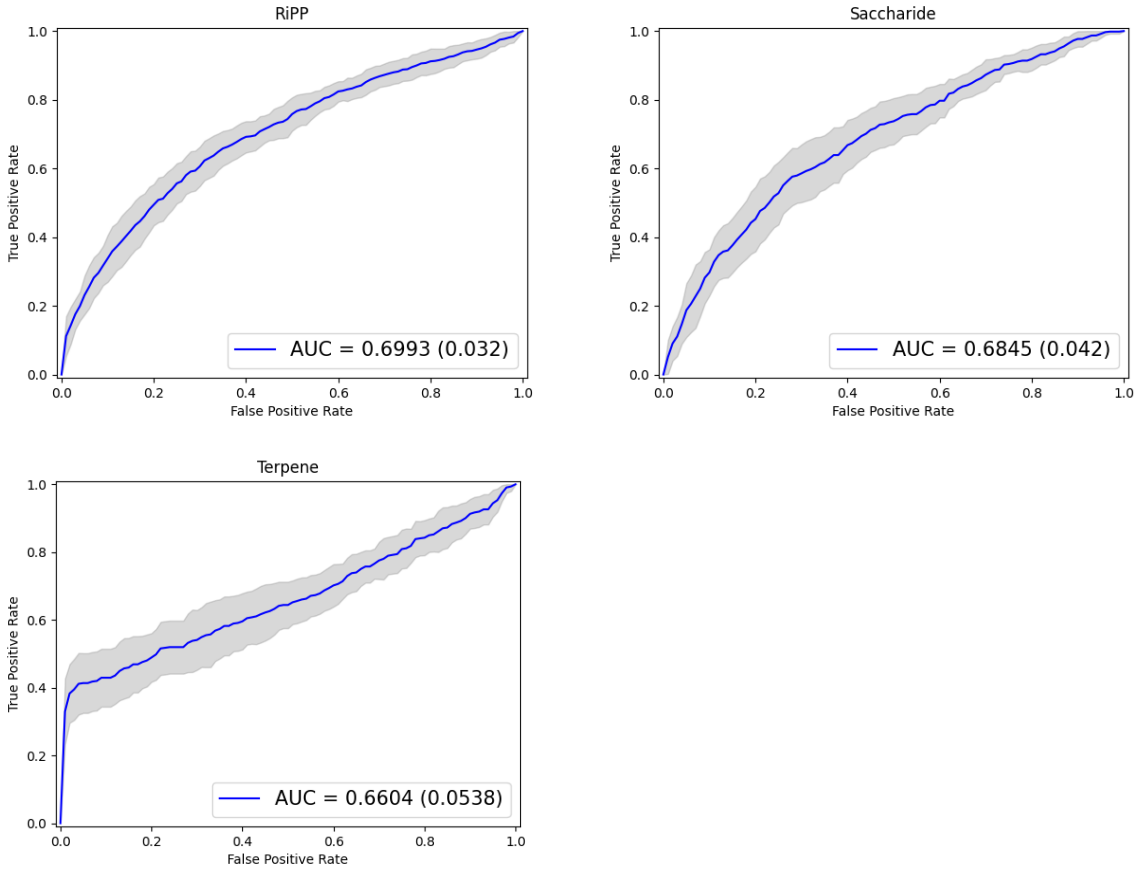


Table 2: AUC of individual labels - best scoring model NN2

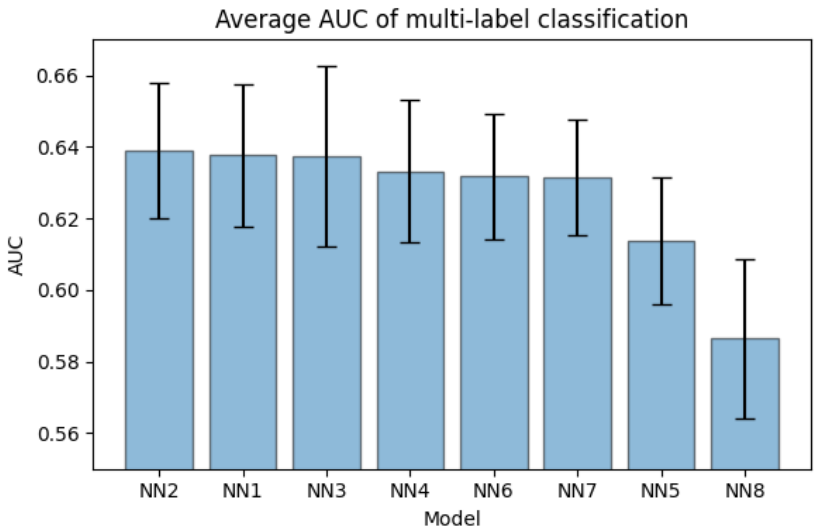


Figure 1: AUC scores of deep learning models

Biosynthetic Gene Clusters classification

Lucie Cervenkova

CTU - FIT

cervel10@fit.cvut.cz

December 2022

Model	Encoding	AUC	Accuracy	F1 Score	Precision	Recall
MLP Classifier (sklearn)	Pfam2Vec	0,9266	0,6958	0,7963	0,8585	0,7578
MLP Classifier (sklearn)	One-Hot Encoding	0,9204	0,7258	0,8287	0,8633	0,8037
Extra Trees (Ensemble)	One-Hot Encoding	0,9146	0,7323	0,8268	0,9005	0,786
Random Forest	Pfam2Vec	0,8838	0,5904	0,7321	0,8904	0,6511
K-Nearest Neighbors	One-Hot Encoding	0,8816	0,7123	0,7926	0,8609	0,7781
Extra Trees (Ensemble)	Pfam2Vec	0,8806	0,5924	0,7351	0,8936	0,6555
K-Nearest Neighbors	Pfam2Vec	0,8711	0,6611	0,765	0,8213	0,7519
Decision Tree	One-Hot Encoding	0,834	0,741	0,8213	0,8262	0,819
Extra Tree	One-Hot Encoding	0,7977	0,6681	0,7645	0,7724	0,7613
Decision Tree	Pfam2Vec	0,7429	0,5583	0,6845	0,6813	0,6911
Extra Tree	Pfam2Vec	0,7229	0,5213	0,6493	0,6478	0,6537
NN2 (2 hidden layers, 100 nodes)	Pfam2Vec	0,639	0,1464	0,2284	0,3737	0,1843
NN1 (1 hidden layer, 100 nodes)	Pfam2Vec	0,6377	0,103	0,1828	0,3802	0,1373
NN3 (3 hidden layers, 100 nodes)	Pfam2Vec	0,6373	0,1615	0,2415	0,3728	0,2037
NN4 (4 hidden layers, 100 nodes)	Pfam2Vec	0,6332	0,1859	0,2665	0,3793	0,2309
NN6 (2 hidden layers, 50 nodes)	Pfam2Vec	0,6318	0,0988	0,1609	0,3284	0,1277
NN7 (1 hidden layer, 50 nodes)	Pfam2Vec	0,6315	0,0699	0,1293	0,2802	0,0947
Random Forest	One-Hot Encoding	0,6254	0,1892	0,3114	0,4952	0,248
NN5 (3 hidden layers, 50 nodes)	Pfam2Vec	0,6139	0,0681	0,1167	0,2019	0,0931
NN8 (1 hidden layer, 10 nodes)	Pfam2Vec	0,5864	0,0043	0,0103	0,0813	0,0059

Table 3: Summary results of all models

Biosynthetic Gene Clusters classification

Lucie Cervenkova

CTU - FIT

cervel10@fit.cvut.cz

December 2022

References

Medema M. *et al.* antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences : Nucleic Acids Research, 2011.

Cimermanic, P. *et al.* Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters: Cell, 2014.

Kautsar, S. *et al.* MIBiG 2.0: a repository for biosynthetic gene clusters of known function: Nucleic Acids Research, 2020.

Hannigan, G. D. *et al.* A deep learning genome-mining strategy for biosynthetic gene cluster prediction: Nucleic Acids Research, 2019. - Vol. 47.