Final Report: Novel antimicrobials for swine health

A Computational Drug Discovery project supported by Ontario Pork and NSERC CRD

Submitted by Chris G. Gray

Department of Physics

University of Guelph

Introduction:

We aimed at exploring the extent to which publicly available data on antimicrobial peptides (AMPs) can be utilized using the state of the art models and training algorithms in machine learning (ML) to yield predictors that can screen any peptide sequence for their antimicrobial activity. Within this project we collected datasets on some pathogens of interest to the pork industry, performed ML trainings on best of the available models for this purpose, optimized the design (hyperparameters) of these models and explored the limits of the training using the currently available data.

Objectives: (original objectives from project proposal)

Enhance the performance and reduce the toxicity of a novel synthetic AMP (HHC-36)

Employ machine learning (ML) methods for discovering new, more potent antimicrobial peptides

Determine the hemolytic activity of these AMPs.

Note: Due to reduction of the grant amount from the proposed 80k to 40k, and following advice from Ontario Pork, all experimental components of this project were postponed to a future phase of the study, in favor of focusing solely on computational studies.

Materials and Methods:

After examining several online datasets, we used *Database of Antimicrobial Activity and Structure of Peptides* (DBAASP) as our major source of data, and the implementation of graph convolutional networks in the deepchem library as our primary learning model and training algorithm. Several training runs were performed for tuning the network parameters, followed by production (training) runs. All of the analysis was performed using suitable python libraries.

Results and Discussion:

We determined the asymptotic limits of the training scores for the graph convolutional models we employed on the available data. Within a mostly uncharted territory, these training results set one of the very first machine learning results on predicting antimicrobial activity of AMPs. What is more, our results show a clear correlation between the dataset size and the final training score.

Conclusions:

Due to reduction of the budget for this project from the requested $80k to $40k, we were not able to extend public datasets by producing more experimental data. Hence, we could not go further than what public data allowed for training our models. We have demonstrated the usefulness of applying graph convolutional models, used almost exclusively on small molecule datasets, to the domain of biologics (peptides and proteins). Our results showed meaningful relationship between the amount of data and the quality of predictors that can be trained, and strongly suggest that investing in dataset development in the domain of biologics is worthwhile and meaningful.

Introduction:

This project focused primarily on employing molecular machine learning (MML) methods to the existing public data on antimicrobial activity of studied antimicrobial peptides (AMPs) from either natural or synthetic origin, to establish an association between chemical structure and activity (or effectiveness). The goal has been to train a model for each dataset such that it can predict antimicrobial activity, given the chemical structure (or alternatively, the amino acid sequence of the peptide). The research was thus divided between data mining efforts to collect and curate datasets on antimicrobial activity (as well as other relevant characteristics such as hemolytic activity (a measure of toxicity)), and adapting the best of machine learning models and training algorithms to utilize the collected data.

Objectives:

Enhance the performance and reduce the toxicity of a novel synthetic AMP (HHC-36)

Employ machine learning (ML) methods for discovering new, more potent antimicrobial peptides

Determine the hemolytic activity (toxicity) of these AMPs.

Materials and Methods:

Several sources of data and several learning algorithms were considered and thoroughly examined during this project.

After our initial efforts on collect the required data from the *Data Repository of Antimicrobial Peptides* (DRAMP, available at http://dramp.cpu-bioinfor.org/), we switched to the *Database of Antimicrobial Activity and Structure of Peptides* (DBAASP, available at https://dbaasp.org) for almost all our data needs. Altogether, five separate datasets were collected, as shown in table 1.

Compared to other AMP databases, DBAASP has a more comprehensive and up-to-date data, is more carefully curated and also provides a basic application programming interface (API), which we used for raw data extraction. Each of the labeled datasets was further manually curated for consistency and cleanliness. We were interested in minimum inhibitory concentration (MIC) of each AMP against pathogens of interest, and we used such values in molar concentrations, as it is a better indicator of per molecule efficacy of an AMP against the pathogen. In cases where concentration were reported in mass units, we made the conversion to molar units, using utility functions provided by the RDKit library (rdkit.org). We also used RDKit tools to convert the amino acid (FASTA) representation of the AMPs to the simplified molecular-input line-entry system (SMILES) representation, suitable for our learning algorithms.

On the algorithmic side, we considered a variety of methods, ranging from the random forest to the most recent graph-based models (1,2). We ultimately decided to use graph convolutional neural networks (GCNNs), as implemented in the DeepChem library (on the web at deepchem.io) as our primary learning model. Choice of GCNNs over other models (including models based on natural language processing models) was based on several recent reports, wherein GCNNs are shown to outperform other models in molecular machine learning tasks. Our choice of deepchem for implementation, was made based on active development and the lively community built around this library, continuous implementation of most successful models within deepchem, as well as very active research on cutting edge MML models by the developers of deepchem at Stanford University.

Training of the GCNNs was performed on computational clusters of Compute Canada, in particular, Graham, Cedar and Niagara. Hyperparameter tuning was performed at the initial stages of the project, and also towards the final stages, when other limiting factors in generalization power of the network were fully investigated.

Activity	Dataset size (# of sequences)	Usage
MIC against salmonella	960	Supervised learning
MIC against E. Coli	1140	Supervised learning
MIC against Listeria Monocytogenes	333	Supervised learning
Hemolytic activity (50% hemolysis)	383	Supervised learning
None (all available sequences)	10,666	Unsupervised learning, semi-supervised learning

Table 1 Summary of the datasets collected and used in this project.

Results and Discussion:

Training a machine learning model is typically done in epochs, with each epoch itself being divided into batches. During each batch, a subset of the data is used to train the model, one step forward to a better state. Quality of a state is assessed by how well the network can perform the task it is expected to accomplish, which in our case is predicting the

MIC value of a given peptide, given its sequence. One epoch consists of as many batch-steps that covers the whole dataset. So in one epoch, all of the available data is visited by the network.

Among all of the training runs that we performed, we show two of the most representative results in figure 1.

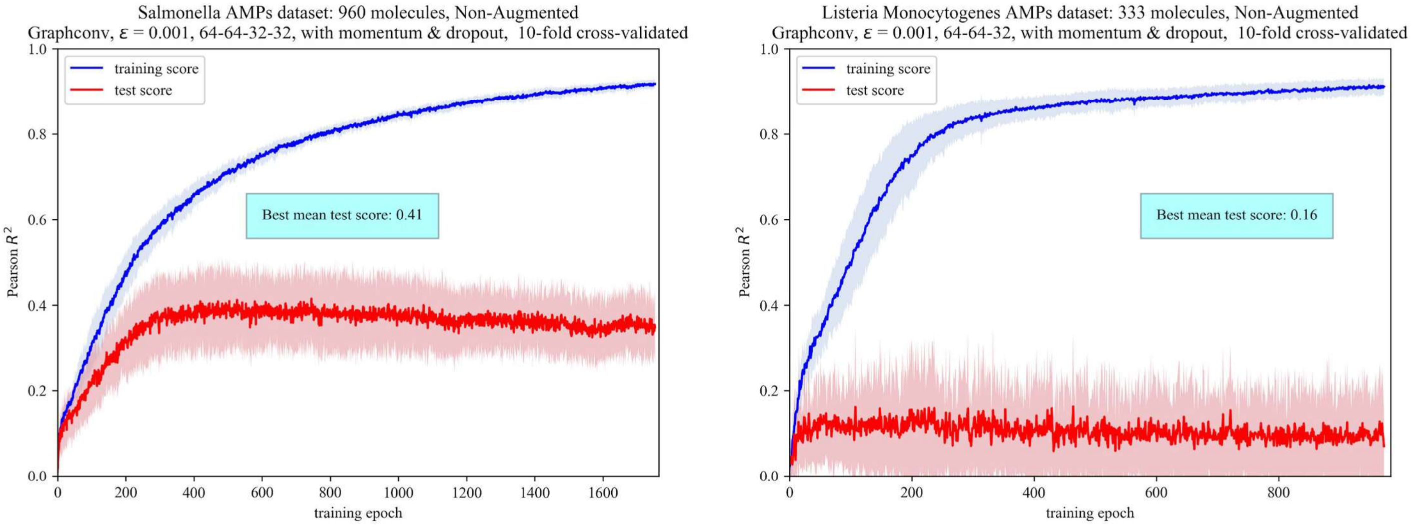

Figure 1: Training curves (with optimal hyperparameters in each case) on the Salmonella (left) and Listeria (right) datasets, with 10-fold cross-validation. Reported best scores are averages among the 10 separate models in each case.

Effect of the dataset size on the best score (1 is ideal) is vivid in figure 1, by comparing between salmonella and listeria datasets, where the ratio of dataset size is very close to 3:1. Test score is seen to very closely follow the ratio of data used for training (the maximum available amount, in each case). We use cross-validation for these runs to minimize the effect of unevenness in the dataset: the dataset is divided into 10 non-overlapping subsets, and 10 training runs are performed, each using one of the 10 subsets as test set (the set used solely for assessment and not for training), and the rest of the data is used for training. The relationship seen here between the test score (predicting power of the trained network) and the dataset size, was typical of all of our training scenarios among all the datasets, network architectures and training settings that we performed.

Aside from the actual machine learning and optimization results, the main outcome of this project is a clear characterization of the extent of data required for a successful AMP discovery and optimization project. The value of this gained insight only becomes clear when the broader landscape of computational drug discovery is considered: most of the reported efforts in this domain focus on small, drug-like molecules. We should distinguish our study from the vast majority of ML-based computational drug discovery research in two ways: first, we focused on anti-microbial peptides, which generally fall under the domain of biologics (peptides and proteins). Characterizing the chemical structure of biologics and then associating it with their properties of interest, is generally a much harder problem compared to performing the same task on small molecules, simply due to higher number of degrees of freedom involved in studying a larger object. Second, we opted for performing a regression, rather than a classification task. This poses extra difficulty, especially in face of data scarcity that we deal with in this domain (3). Simply put, predicting whether a given peptide is or is not showing activity against a given target (e.g., salmonella) is a much simpler task than quantifying the activity. We firmly believe that a learning algorithm can provide useful information/predictions in this domain only if the outputs are quantitative, and not merely categorical. Our choice of performing regression tasks was made with this in mind.

Conclusions:

To the best of our knowledge, this project was one of the few of this nature to have ever been performed on ML-based screening of antimicrobial peptides, and the first ever to use GCNNs on biologics. We later were able to use a more recent, mixed natural language processing (NLP) model (4) on the same data, towards the same regression task. Although the results were improved (R^2 between 0.5 and 0.6 for the salmonella dataset), they are still not reliable enough for commercial use.

The correlation between dataset size and training quality, as well as our later results on the same datasets with a newer NLP-based model, clearly suggest strong algorithmic potential for screening AMPs. This, along with the need for new AMPs, which initially inspired this project, points at the need in investment in dataset development through high-throughput screening of AMPs against various pathogens and therapeutic protein targets.

Acknowledgments:

We wish to express our sincere gratitude to Ontario Pork for making this research possible. We are also deeply thankful to Prof. Graham Taylor, head of the machine learning research group (MLRG) at the University of Guelph, who supervised the machine learning investigations in this project. We also express thanks to University of Guelph professors Alejandro Marangoni, Food Science, Michael Rogers, Food Science, and Robert Friendship, DVM, for valuable help and discussions in the early stages of the project.

References:

1. Kearnes, Steven, et al. "Molecular graph convolutions: moving beyond fingerprints." *Journal of computer-aided molecular design* 30.8 (2016): 595-608.
2. Duvenaud, David K., et al. "Convolutional networks on graphs for learning molecular fingerprints." *Advances in neural information processing systems*. 2015.
3. Imrie, Fergus, et al., "Protein family-specific models using deep neural networks and transfer learning improve virtual screening and highlight the need for more data." *Journal of chemical information and modeling* 58.11 (2018): 2319-2330.
4. Gómez-Bombarelli, Rafael, et al., "Automatic chemical design using a data-driven continuous representation of molecules." *ACS central science* 4.2 (2018): 268-276.

Knowledge Transfer:

Our project was covered in **The Global Shift** section of the *Ontario Hog Farmer* magazine, in an article entitled "*Using Machine Learning in the Search for Antimicrobial Alternatives*", by Lilian Schaer. It is available online at:

http://www.livestockresearch.ca/novel-antimicrobials-alternatives-for-swine/

A news report of the project by Jackie Clark "Finding Alternatives to Pig Antibiotics" was published in the February 10 2020 issue of Farms.com Newsletter. It is available online at : http: // www.farms.com/cg-industry-news/finding-alternatives - to pig-antibiotics-668.aspx.