Robust Statistical Methods for Toxin Prediction

1 Introduction

Recent advances in synthetic biology have led to very powerful techniques for design and assembly of gene sequences, circuits and metabolic pathways [3, 5]. This has significant implications for domains such as biotechnology, agriculture, drug discovery, and healthcare. However, this also increases the accessibility of toxic protein sequences either accidentally or adversarially. As a result, there is strong interest in developing accurate machine learning methods for predicting whether a given sequence is a toxin. There have been two broad directions in prior work on toxin prediction. The first involves using sequence based features, e.g., ToxinPred [2] and BTXpred [6] both use a Support Vector Machine (SVM) approach to predict toxins, using sequence motifs as predictors. Both are trained on fairly small datasets, with fewer than 1000 toxin sequences. Jain and Kihara [4] design a neural network method using gene ontology (GO) terms as predictors, instead of sequence motifs. They report an accuracy of 0.877 for toxins.

In this paper, we study to what extent additional attributes can improve toxin prediction, beyond the performance of [4] using GO terms. Further, we are also interested in more specialized toxin categories, e.g., neurotoxins and ion channel toxins. Given that neural network approaches do not give very interpretable results, we also explore whether simpler and more interpretable models can give comparable or better performance. Our contributions are the following:

1. Improving toxin prediction using LASSO. We show that the Least Absolute Shrinkage and Selection Operator (LASSO) regression for feature selection [7] has a Positive Predictive Value (PPV) value of over 0.98 for toxins, which is significantly better than the approach of [4] for toxin prediction. We are the first to consider prediction of sub-classes of toxins, and Lasso has a PPV of 0.90 and 0.87 for neurotoxin and ion-channel toxin, respectively. However, one limitation is that some of the attributes selected are too specific (e.g., particular taxons), which potentially limit the robustness of the method. This motivates our next result.

2. Improving robustness using Association Regularized Classification Trees (ARCT). We introduce a novel data exploration algorithm based on a sequence of models which is guided by the hierarchical structure of the sequence data. We first construct an association tree from the data, and use it to fit a sequence of classification trees by multiple passes over the data. We demonstrate this model by classifying sequences as toxic or non-toxic and show that it produces explainable models with good positive predictive performance. We find that within each pass, there is a general trend of PPV increasing while recall decreases as the models progress. This approach yields interesting relationships in specific regions of the feature space while retaining good predictive capability.

Details of our methodology and our results are omitted due to the space limit. A complete version of our paper is available at https://github.com/NSSAC/MLCB2019-toxin-prediction/blob/master/Toxin_Prediction_MLCB_2019.pdf.

2 Methodology

2.1 Multinomial Logistic Regression with LASSO penalty

We use multinomial logistic regression for multi-class toxin prediction. Since the number of attributes is very large, the LASSO penalty is implemented to obtain a sparse model. Multinomial logistic regression is a generalization of logistic regression when there are more than two outcomes. Let $y_i \in \{1, \ldots, K\}$ denote the i^{th} response, for $i = 1, 2, \ldots, n$. Let $\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots, x_{iM})$ denote the vector of attributes for the i^{th} response. Then we have

$$Pr(y_i = k) = \frac{e^{\mathbf{x_i}\beta_k}}{1 + \sum_{h=2}^{K} e^{\mathbf{x_i}\beta_h}}, k = 0, \ldots, K,$$

where $\beta_j = (\beta_{j0}, \beta_{j1}, \beta_{j2}, \ldots, \beta_{jM})$, and β_{jm} is the regression coefficient associated with the m^{th} attribute and the k^{th} outcome, and β_0 is the all 0-vector. The likelihood is

$$\mathcal{L} = \prod_{i=1}^{n} \Big[\frac{1}{1+\sum_{h=2}^{K} e^{\mathbf{x_i}\mathbf{fi}_h}}\Big]^{I(y_i=1)} \cdots \Big[\frac{e^{\mathbf{x_i}\mathbf{fi}_k}}{1+\sum_{h=2}^{K} e^{\mathbf{x_i}\mathbf{fi}_h}}\Big]^{I(y_i=k)} \cdots \Big[\frac{e^{\mathbf{x_i}\mathbf{fi}_K}}{1+\sum_{h=2}^{K} e^{\mathbf{x_i}\mathbf{fi}_h}}\Big]^{I(y_i=K)},$$

where $I(y_i = k) = 1$ if $y_i = k$, and $I(y_i = k) = 0$ if $y_i \neq k$.

The regression coefficients are estimated by minimizing the negative penalized loglikelihood below,

$$\hat{\beta} = \text{argmin}_\beta \Big\{ -\frac{1}{n}\text{Log}\mathcal{L} + \lambda Pen(\beta) \Big\},$$

where λ is the tuning parameter, and $Pen(\beta) = ||\beta||_1$ is the LASSO penalty. In our study, the tuning parameter λ is chosen by minimizing the 5-fold cross-validation error.

2.2 Association Regularized Classification Trees (ARCT)

A downside to applying the LASSO to the toxin data is the presence of highly predictive yet uninteresting keywords. One example is the keyword "KW-0181", which indicates the presence of a complete proteome. Relative to the test and training set, this is an excellent predictor, but biologically it is uninteresting. Further examples can be seen in the list of keywords manually removed from the LASSO training data. Some of these are so broad as to be meaningless, whereas some of them are nearly identical to the thing we wish to predict.

While is it impossible to quantify "interest" in an objective way, Association Regularized Classification Trees (ARCTs) give us a way to automatically filter overly broad or tautologically predictive terms in a way that doesn't compromise predictive power.

2.2.1 Association Tree

In this paper, an association tree is a directed acyclic graph (DAG) over the training data features. To create this graph, we form a directed edge from term A to term B, $A \to B$, whenever the presence of B implies the presence of A:

$$P(A = 1|B = 1) = 1,$$

where P means empirical proportion. This form of association reconstructs hierarchies within biological classification systems, as well discovers new links between them.

If it happens that we form both $A \to B$ and $B \to A$, we combine both terms into a single term and concatenate their names. This can happen, for instance, if one had a data set where all sequences came from humans. The terms *homo* and *sapiens* would appear equivalent, even though this is not the case generally. After all such sets of terms are reduced, the resulting graph is a DAG.

The terms in this DAG are classified into layers based on their parent terms. The first layer is all terms with no parent terms. The second layer consists of all terms with no parents *outside of the first layer*. In general, the n^{th} layer is all terms with no parents outside of the first $n-1$ layers. Terms in the first layers are broader and appear in a greater proportion of sequences, while terms in later layers correspond to fewer sequences and denote finer structures and attributes.

2.2.2 Classification Tree Sequences

We use this association tree to fit a sequence of classification trees. We proceed by selecting all terms in the first layer and fitting a cross validated regression tree. Next, we take the terms the tree used to classify the sequences and collect all child terms of those terms in the second layer, again fitting a tree model. This procedure continues either until there are no remaining terms with sufficient classification power, or the terms that are split on have no children. This sequence forms the first *pass* of the algorithm.

Subsequent passes proceed the same way, except that all terms that were used in the first pass are removed from consideration. In theory this process can continue until every term

has been exhausted, but we have found that the performance begins to deteriorate to the point of uselessness after around 10 passes.

Early models in the sequence have optimal predictive performance, being competitive with methods based on the LASSO or neural networks [4]. While performance generally deteriorates as the sequence advances, this loss in predictive capability is made up for by explanatory utility, as the models are based on more specific biological features. In this way, we can find interesting relationships in specific regions of the feature space while retaining strong predictive capability.

2.3 Data Sets

We follow a similar approach as [4] for constructing our data. We extract 6341 protein sequences from the UniProtKB database [1], which are annotated with a Toxin Keyword 800. We further partition the toxin sequences into four sets: neurotoxin, ion channel toxin, neuro and ion channel toxin, and other toxin, based on the associated keywords. In order to have negative examples, we sample 708,839 sequences which are non-toxin. Together, this gives us a dataset of 715,180 sequences.

For each sequence, we identify a diverse set of attributes, which contain features like keywords, gene-ontology (GO) terms, IPR numbers, UniProt taxon ids, and Uniref ids (total of 2,746,614 unique attributes). We remove the following keywords from the attribute set, because they denote obvious toxin information: 'kw-0181', 'kw-0800', 'kw-0260', 'kw-0528', 'kw-0872', 'kw-0959', 'kw-1061', 'kw-1185', 'kw-1199', 'kw-1213', 'kw-1216', 'kw-1217', 'kw-1255', 'kw-0123', 'kw-1254', and 'kw-0843'. For efficiency reasons, we pick a random subset of 40,000 non-toxin sequences for the training dataset. We consider two types of classification problems. The first is a binary toxin/non-toxin prediction, and the second is a multinomial response, which involves predicting one of the following classes: neurotoxin, ion channel toxin, neuro and ion channel toxin, other toxin, or non-toxin.

3 Results

3.1 Multinomial Logistic Regression with LASSO penalty

Table 1 shows the performance of our model on the entire dataset. Additional details, along with the selected attributes are presented in the Supplementary Material.

Table 1: The confusion matrix obtained by using the multinomial logistic regression model in Table 6 to predict all data. The PPV is Positive Predictive Value. The overall accuracy is 0.9990.

		Ground Truth						
		Non-toxin	Other toxin	Neuro	Ion	Neuro & Ion	sum	PPV
Prediction	Other toxin	**708771**	75	3	0	0	708849	0.9999
	Other toxin	68	**3085**	262	28	72	3515	0.9777
	Neuro	0	32	**508**	0	21	561	0.9055
	Ion	0	18	3	**327**	24	372	0.8790
	Neuro & Ion	0	0	20	118	**1745**	1883	0.9267
	sum	708839	3210	796	473	1862	715180	
	Recall	0.9999	0.9611	0.6382	0.6913	0.9372		

3.2 ARCT Results

We trained the ARCT model on a subset of 150,000 sequences and validated on a hold out set of 50,000. We fit a total of 43 trees over 7 sweeps along the data. PPV and recall for each tree is provided in the supplemental, in Figure 3. Within each pass, there is a general trend of PPV increasing while recall decreases as the models progress. A similar pattern holds as one moves through the passes, though it is less pronounced. Once the pass count gets high enough, performance collapses along both dimensions.

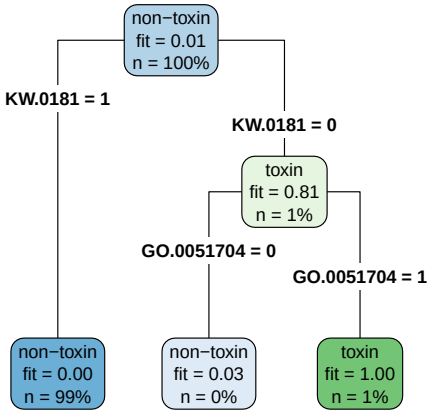

Figure 1: The first discovered tree. This tree splits on very general terms

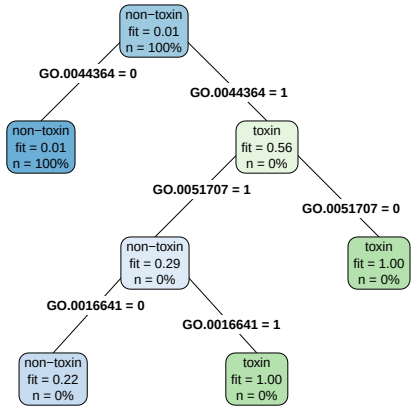

Figure 2: A deeper tree splitting on GO terms

Figures 1 and 4 in the supplement show this pattern. Figure 1 is the first tree fit on the data. The terms "KW-0818" and "GO:0051704" refer to sequences which are part of a proteome or which have to do with the interaction between proteomes. Evidently, toxicity is associated with the absence of both of these properties. ANDREW WHYYYYYYY?!

Contrast with Figure 4, which splits on taxon classifiers. These are much more specific terms, referring to individual species or small groups of them. This tree has poorer performance than the first tree, but provides a more specific explanation. Interestingly, many of the trees in the 6^{th} layer are comprised primarily of taxon IDs and classify on distinct groups of species. The tree in Figure 4 in the appendix contains IDs for species in the clade of protostomia. This tree's predictive performance is middling, with a PPV of 0.87 and a recall of 0.54, but it's simple structure serves as an explanation: this tree is classifying on sub-taxa of protostomia which it found to be toxic.

Table 2: GO term descriptions.

Term ID	Description
0044364	Disruption of cells of other organism
0051707	Response to other organisms
0016641	Oxidoreductase activity, acting on the CH-NH2 group of donors

A fascinating example is shown in Figure 2, which splits on a sequence of GO terms. Descriptions of the relevant terms are given in Table 2. In words, this tree will classify a sequence as toxic if it involves disruption of the cells of other organisms and is not a response to other organisms. If it is both, it will classify as toxic if in addition it is involved in oxidoreductase activity. This tree has a PPV of 0.95 and a recall of 0.08, so while it is a very poor rule for finding toxins generally, it is very reliable in regards to those sequences which it does classify as toxic. This means that this tree is identifying a particular *class* of toxin, as well as hinting at possible mechanisms of action for that toxin.

4 Conclusion

We find that the LASSO model gives good predictive power for toxins and sub-categories of toxins, when an expanded set of attributes (beyond GO terms) are used. However, only a small fraction of data on toxins in the real world is available in the UniProt database, and so the very specific predictors used by this method might limit performance. We propose a new ARCT model, which alleviates this problem by identifying a hierarchical structure and using it to train a sequence of classification trees, which provide a tradeoff between PPV and recall.

References

[1] T. U. Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, 11 2018.

[2] S. Gupta, P. Kapoor, K. Chaudhary, A. Gautam, R. Kumar, and G. P. Raghava. In silico approach for predicting toxicity of peptides and proteins. *PloS one*, 2013.

[3] R. Hughes and A. Ellington. Synthetic dna synthesis and assembly: Putting the synthetic in synthetic biology. *Cold Spring Harbor Perspectives in Biology*, 9:a023812, 01 2017.

[4] A. Jain and D. Kihara. Gene ontology-based protein toxicity prediction using deep learning. In *ASM Biothreats*, 2019.

[5] C. Plesa, A. M. Sidore, N. B. Lubock, D. Zhang, and S. Kosuri. Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. *Science*, 359(6373):343–347, 2018.

[6] S. Saha and G. Raghava. Btxpred: prediction of bacterial toxins. *In Silico Biol*, page 405-12, 2007.

[7] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, page 267–288, 1996.

Supplementary Material

5 Details about the dataset

We train our models on the downsampled data and predict the downsampled data, then we obtain the in-sample results in Tables 6 and 3. Also, we use the models to predict the whole data, then we obtain the results in Tables 8 and 5. In addition, we randomly partition the downsampled data into a training set and a test set in an 80/20 split. We train the model on the training set and predict the test set, so that we have results in Tables 7 and 4.

6 Additional results using the LASSO method

6.1 Toxin prediction

Tables 3, 4 and 5 show the results of logistic regression for the binary toxin prediction problem.

Table 3: The in-sample confusion matrix for the binary toxin prediction. The PPV is Positive Predictive Value. The overall accuracy is 0.9983.

		Ground Truth			
		Non-toxin	Toxin	sum	PPV
Prediction	Non-toxin	**39997**	75	40075	0.9981
	Toxin	3	**6266**	6269	0.9995
	sum	40000	6341	46341	
	Recall	0.9999	0.9882		

Table 4: The 80/20 out-of-sample confusion matrix for the binary toxin prediction. The PPV is Positive Predictive Value. The overall accuracy is 0.9988.

		Ground Truth			
		Non-toxin	Toxin	sum	PPV
Prediction	Non-toxin	**7966**	10	7976	0.9987
	Toxin	1	**1291**	1292	0.9992
	sum	7967	1301	9268	
	Recall	0.9999	0.9923		

Table 5: The confusion matrix obtained by using the logistic regression model in Table 3 to predict all data. The PPV is Positive Predictive Value. The overall accuracy is 0.9998.

		Ground Truth			
		Non-toxin	Toxin	sum	PPV
Prediction	Non-toxin	**708774**	75	708849	0.9999
	Toxin	68	**6263**	6331	0.9893
	sum	708842	6338	715180	
	Recall	0.9999	0.9882		

6.2 Multinomial Logistic Regression with LASSO penalty

Tables 6 to 8 show the results of multi-toxin classification, and Table 9 summarizes the selected attributes. Table 6 shows the in-sample performance, Table 7 shows the 80/20 out-of-sample performance, and Table 8 shows the performance using the in-sample model to predict for the whole data. Tables 3 to 5 show the results of binary toxin classification, and Table 10 summarizes the selected attributes. Table 3 shows the in-sample performance, Table 4 shows the 80/20 out-of-sample performance, and Table 5 shows the performance using the in-sample model to predict for the whole data.

Table 6: The in-sample confusion matrix for the multi-toxin prediction. The PPV is Positive Predictive Value. The overall accuracy is 0.9853.

		Ground Truth						
		Non-toxin	Other toxin	Neuro	Ion	Neuro & Ion	sum	PPV
Prediction	Other toxin	**39997**	75	3	0	0	40075	0.9981
	Other toxin	3	**3085**	262	28	72	3450	0.8942
	Neuro	0	32	**508**	0	21	561	0.9055
	Ion	0	18	3	**327**	24	372	0.8790
	Neuro & Ion	0	0	20	118	**1745**	1883	0.9267
	sum	40000	3210	796	473	1862	46341	
	Recall	0.9999	0.9611	0.6382	0.6913	0.9372		

Table 7: The 80/20 out-of-sample confusion matrix for the multi-toxin prediction. The PPV is Positive Predictive Value. The overall accuracy is 0.9861.

		Ground Truth						
		Non-toxin	Other toxin	Neuro	Ion	Neuro & Ion	sum	PPV
Prediction	Other toxin	**8005**	18	0	0	0	8023	0.9978
	Other toxin	0	**621**	43	2	15	681	0.9119
	Neuro	0	8	**114**	0	6	128	0.8906
	Ion	0	1	0	**53**	8	62	0.8548
	Neuro & Ion	0	0	1	27	**346**	374	0.9251
	sum	8005	648	158	82	375	715180	
	Recall	1.0000	0.9583	0.7215	0.6463	0.9227		

Table 8: The confusion matrix obtained by using the multinomial logistic regression model in Table 6 to predict all data. The PPV is Positive Predictive Value. The overall accuracy is 0.9990.

		Ground Truth						
		Non-toxin	Other toxin	Neuro	Ion	Neuro & Ion	sum	PPV
Prediction	Other toxin	**708771**	75	3	0	0	708849	0.9999
	Other toxin	68	**3085**	262	28	72	3515	0.9777
	Neuro	0	32	**508**	0	21	561	0.9055
	Ion	0	18	3	**327**	24	372	0.8790
	Neuro & Ion	0	0	20	118	**1745**	1883	0.9267
	sum	708839	3210	796	473	1862	715180	
	Recall	0.9999	0.9611	0.6382	0.6913	0.9372		

Table 9: Names of selected attributes.

	1	2	3	4	5	6
1	go:0008200	go:0009405	go:0044419	ipr036574	kw-0738	taxon:259437
2	taxon:6855	taxon:6856	taxon:70336	go:0004623	go:0072556	taxon:8689
3	kw-0632	taxon:6489	taxon:69555	ipr023355	kw-0645	go:0019835
4	taxon:6894	kw-0960	kw-0108	go:0035792	kw-0008	kw-1265
5	taxon:6917	kw-1214	taxon:209901	uniref90__d2y204	uniref50__p01414	ipr002223
6	kw-0722	ipr003582	kw-1028	taxon:117992	ipr012499	uniref50__q5y4x5
7	uniref90__d2y240	taxon:115405	taxon:115407	taxon:41361	taxon:41364	taxon:61985
8	taxon:7540	go:0051609	ipr000395	ipr012928	ipr036248	taxon:561
9	taxon:562	uniref50__p83234	ipr013871	kw-1275	taxon:278059	taxon:278060
10	taxon:1340129	taxon:89438	taxon:319920	taxon:1340135	taxon:101317	taxon:128512
11	kw-0102	taxon:89451	taxon:1340065	taxon:1736779	uniref50__a6yr33	uniref50__q9bpe1
12	uniref50__d2y262	taxon:1295017	taxon:1295018	taxon:230230	taxon:93698	ipr015882
13	pog091h0ijt	taxon:6125	taxon:356378	taxon:33318	taxon:33319	taxon:37848
14	taxon:406442	taxon:406443	uniref50__p0dn06	ipr008430	ipr037040	uniref50__q92h62
15	ipr006891	uniref50__q5k5r0				

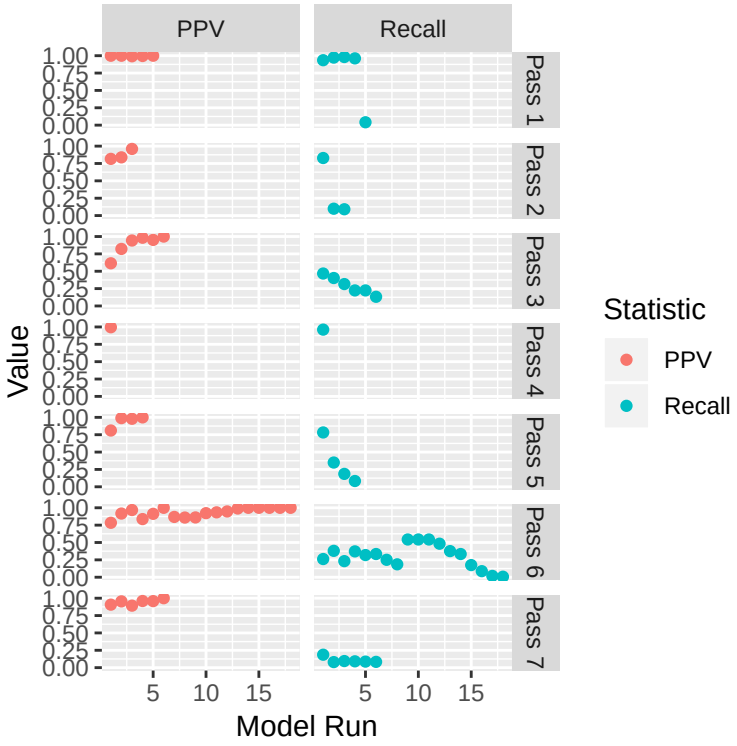

Figure 3: Positive predictive values (PPV) and recalls for the first six passes and the models therein.

Table 10: Names of selected attributes.

	1	2	3	4	5	6
1	go:0009405	go:0044419	kw-0964	ipr027582	taxon:41954	taxon:561
2	taxon:562	ipr015882	pog091h0ijt	ipr008430	ipr037040	uniref50_q92h62
3	ipr006891	uniref50_q5k5r0				

7 ARCT Supplemental Material

Figure 3 shows the PPV and recall for the first six passes in the ARCT sequence. Rows correspond to passes, first at the top, and the points within them to models within each pass. Within each pass, there is a general trend of PPV increasing while recall decreases as the models progress. A similar pattern holds as one moves through the passes, though it is less pronounced. Once the pass count gets high enough, performance collapses along both dimensions. None of the passes beyond what is shown achieve a PPV or recall that exceeds 0.5.

The pattern is sensible if one thinks of these models as "moving down" the term tree (visualizing the root as being at the top). Splits near the root are quite coarse, correctly discovering more toxic sequences while also producing more false positives. Trees on more distal terms are more focused, finding structure in smaller regions of the overall tree while missing toxins characterized by terms in distant branches.

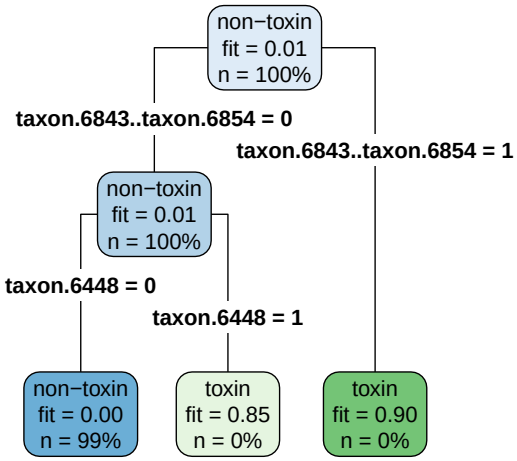

Figure 4: A deeper tree, which has discovered structure based on the taxon IDs of the sequences. Note that taxa 6843 and 6854 are equivalent in these data.