SIGNAL PEPTIDE CLASSIFIERS: A REPORT

Authors: Joshua Koopmans, Michelle de Groot, Thijs Weenink

Signal peptides in peptide sequences are able to be identified. This can be done manually but by automating this process it could potentially be much easier and even more accurate. Several algorithms exist for the classification of data. A selection of classifiers was tested to find the model with the highest accuracy for identifying signal peptides.

Support Vector Machine

The support vector machine (SVM) attempts to find the most optimal line to discriminate data point into classes. The output of the linear function is used to classify; if the output is greater than one, the data point is assigned to one class, and if the output equals minus one, a data point is assigned to the other class.

Logistic Regression

While linear models are often used for predictions, the logistic model is used for classification. Logistic regression algorithms also use a linear function with outputs ranging from negative infinity to positive infinity. To be able to classify, the output should give one of two values; yes (1) or no (0). This is accomplished by “squashing” the outputs of the linear model into a range of 0 to 1 using the sigmoid function.

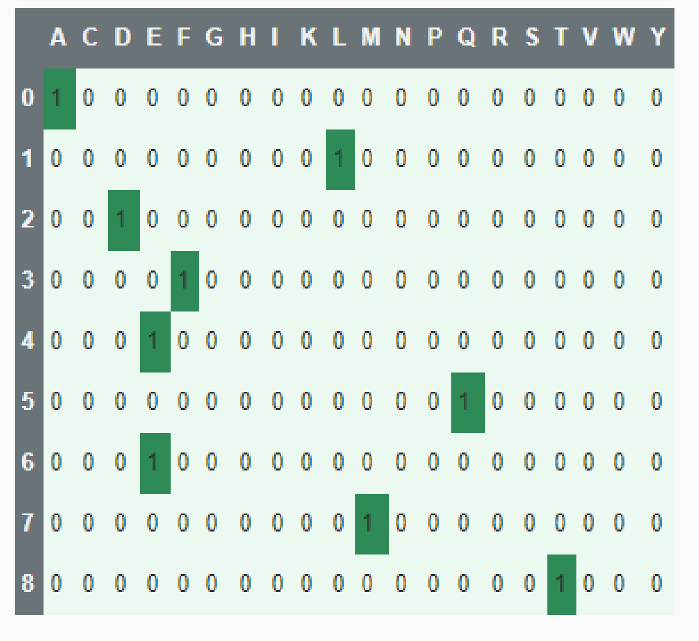
Random Forest

Another classification model is random forest. Random forest is a collection of decision trees with every tree giving a predicted output. This method looks for features to discriminate on. The tree with the highest amount of votes becomes the prediction model.

For the computer to understand our data, we have to convert our non-machine-readable data into machine readable data. For this we can use encoders.

One-hot encoding

One of the simplest encoders is the one-hot encoder. Here, you have a 20-column vector, one column for every amino acid, with for every amino acid in your input sequence a 1 where the amino acids match (**Figure 1**).



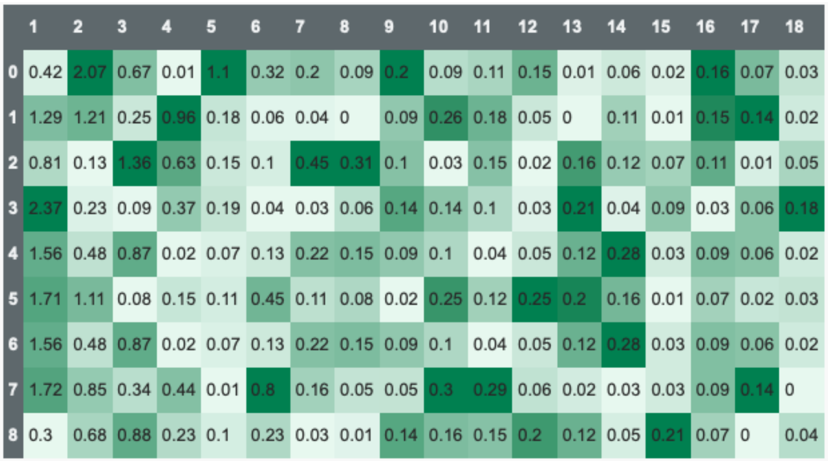
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
6	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0

Figure 1: *Representation of the one-hot encoding.*

Using the one-hot encoder, the algorithms described above were tested. Taking the disbalanced input data into account and using the validation set, the logistic regression algorithm (0.973) outperformed SVC (0.965) and random forest (0.967) when looking at the Area Under the ROC Curve (AUC) score. After using the test data, which has never been seen before by the classifier, both the linear regression and SVC algorithms yielded an AUC score of 0.989 while random forest achieved the best performance with an AUC score of 0.992. See **Table 1** for an overview of the AUC scores per classifier per encoder.

NLF encoding

Developed by Nanni and Lumini in 2009, this method takes the physiochemical properties of amino acids and transforms them into a reduced set of features able to describe amino acids (**Figure 2**).



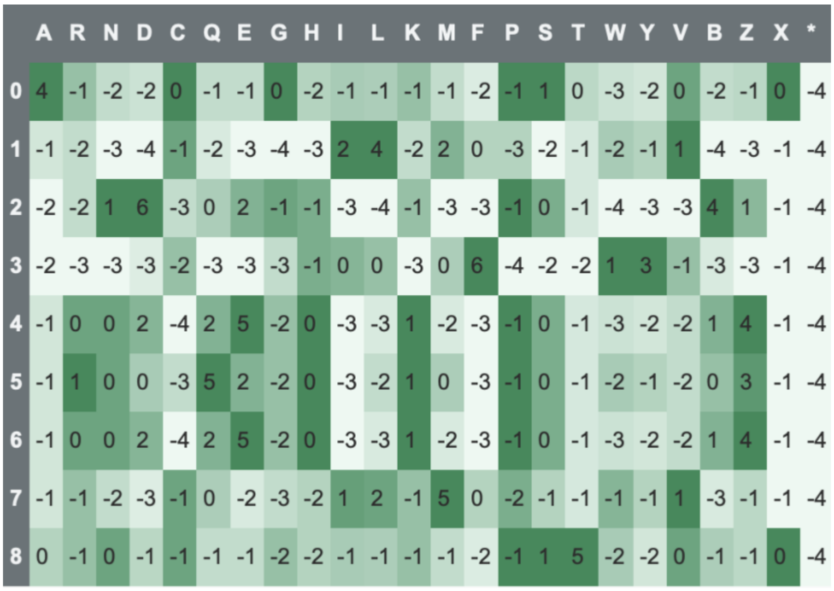
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
0	0.42	2.07	0.67	0.01	1.1	0.32	0.2	0.09	0.2	0.09	0.11	0.15	0.01	0.06	0.02	0.16	0.07	0.03
1	1.29	1.21	0.25	0.96	0.18	0.06	0.04	0	0.09	0.26	0.18	0.05	0	0.11	0.01	0.15	0.14	0.02
2	0.81	0.13	1.36	0.63	0.15	0.1	0.45	0.31	0.1	0.03	0.15	0.02	0.16	0.12	0.07	0.11	0.01	0.05
3	2.37	0.23	0.09	0.37	0.19	0.04	0.03	0.06	0.14	0.14	0.1	0.03	0.21	0.04	0.09	0.03	0.06	0.18
4	1.56	0.48	0.87	0.02	0.07	0.13	0.22	0.15	0.09	0.1	0.04	0.05	0.12	0.28	0.03	0.09	0.06	0.02
5	1.71	1.11	0.08	0.15	0.11	0.45	0.11	0.08	0.02	0.25	0.12	0.25	0.2	0.16	0.01	0.07	0.02	0.03
6	1.56	0.48	0.87	0.02	0.07	0.13	0.22	0.15	0.09	0.1	0.04	0.05	0.12	0.28	0.03	0.09	0.06	0.02
7	1.72	0.85	0.34	0.44	0.01	0.8	0.16	0.05	0.05	0.3	0.29	0.06	0.02	0.03	0.03	0.09	0.14	0
8	0.3	0.68	0.88	0.23	0.1	0.23	0.03	0.01	0.14	0.16	0.15	0.2	0.12	0.05	0.21	0.07	0	0.04

Figure 2: *Representation of the NLF encoding.*

Using the NLF encoder, the algorithms described above were once again tested. Taking the disbalanced input data into account, logistic regression (0.975) and SVC (0.974) outperformed random forest (0.967) when looking at the Area Under the ROC Curve (AUC) score. After testing on the test set, unseen by the classifier, logistic regression has a shown maximum AUC score of 0.993 while SVC came in second place with a score of 0,992 and random forest having the lowest AUC score of 0.966. See **Table 1** for an overview of the AUC scores per classifier per encoder.

BLOSUM62 encoding

The BLOSUM encoder utilizes a substitution matrix to specify the similarity between amino acids using a score (**Figure 3**). The higher the score the more similar the amino acids. The 62 in BLOSUM62 indicated the identity percentage.



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
0	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
1	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
2	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
3	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
4	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
5	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
6	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
7	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
8	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4

Figure 3: *Representation of the BLOSUM62 encoding.*

Using the BLOSUM encoder, the algorithms described above were tested. Taking the disbalanced input data into account, random forest (0.971) slightly outperformed logistic regression (0.963) and SVC (0.958) when looking at the Area Under the ROC Curve (AUC) score. Using the separate data never seen before by the classifier, the test set, random forest takes the win with an AUC score of 0.990 and logistic regression (0.961) and SVC (0.985) lagging behind. See **Table 1** for an overview of the AUC scores per classifier per encoder.

Table 1: *Classifier AUC scores with their respective encoders. The NLF encoding in combination with logistic regression and SVC show the highest accuracy.*

	One-hot encoding	**NLF encoding**	**BLOSUM62 encoding**
Validation set			
Logistic regression	**0.973**	**0.975**	0.963
SVC	0.965	0.974	0.958
Random forest	0.967	0.967	**0.971**
Test set (unseen by classifier)			
Logistic regression	0.989	**0.993**	0.961
SVC	0.989	0.992	0.985
Random forest	**0.992**	0.966	**0.990**

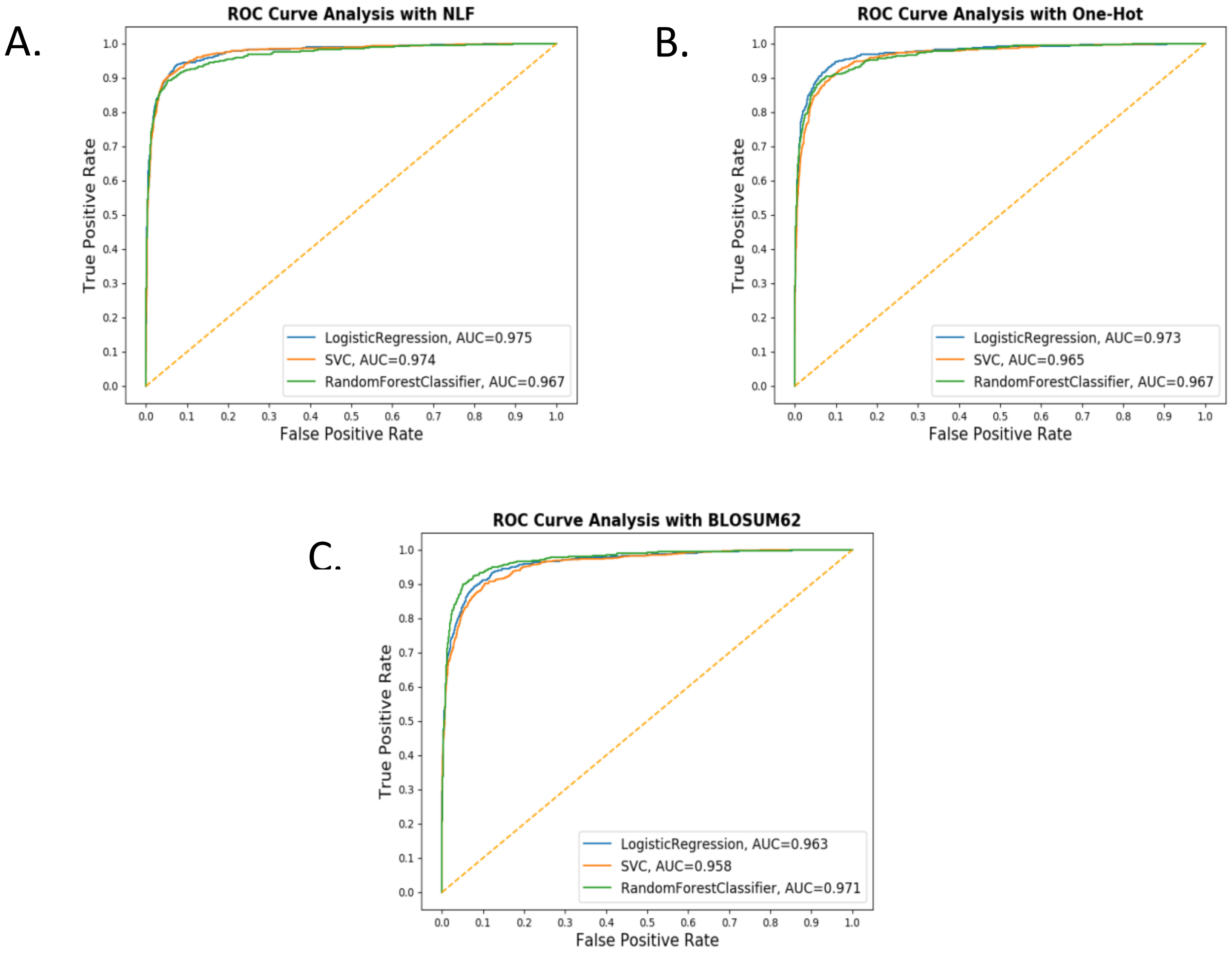


Figure 4: *ROC plots of the classifiers using the validation set (part of the train set).*

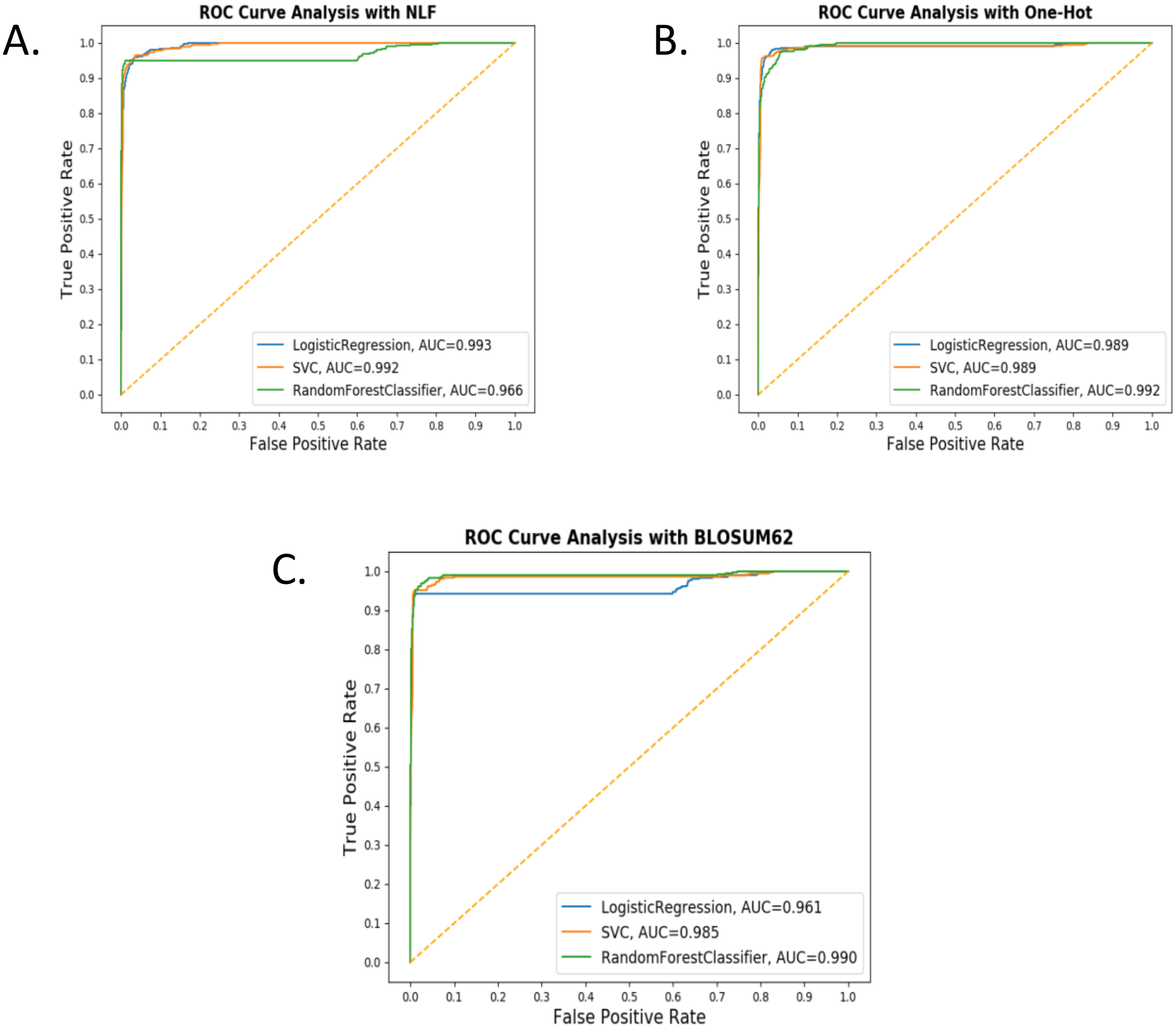


Figure 5: *ROC plots of the classifiers on the test set, never seen before by the classifiers.*

Conclusion

After seeing the preliminary scores, it was clear that the classification of peptide sequences having a signal peptide or not was a linear problem. This is because algorithms like random forest did not yield a better performance compared to the linear algorithms (SVC and logistic regression). Additionally, we noticed the imbalance of our input data; there are far more sequences with the class “with signal peptide” than there are sequences with the class “without signal peptide”. This was rectified by balancing the classes using class weights. Given that the AUC scores are more comprehensive compared to the standard accuracy scores and prove to be better for binary classifications, we opted to use the AUC scores instead of the less meaningful accuracy score. This, even though the accuracy score was higher compared to the AUC scores. Provided the scores were higher while using the test set (which has never been seen before by the classifiers) compared to the validation set (part of the train set), it confirms that the validation set is a good representation of the test set (**Figure 4**; **Figure 5**; **Table 1**). Given that the AUC scores per algorithm is minimal while using the test set (random samples of the true distribution), it is possible that random forest has the best performance by chance due to the test set containing a higher number of samples that are classified better with random forest (**Table 1**). Additionally, it is possible that the test set is not a good representation of the train set, which also contains the validation set. This could be rectified by using an ensemble, which utilizes all parts of the train set. Overall, the simpler one-hot encoding performed the best by all classifiers while the more useful encoder utilizing amino acid properties had a great performance with the linear classifiers. Furthermore, the BLOSUM62 encoder performed better with the non-linear random forest classifier. **Finally, we think it is better to use the NLF encoder due to the combination of a high score and its use of amino acid properties. In addition, we would recommend using logistic regression, a linear classifier.**