A classifier approach to predict protein secretion in *Aspergillus niger*

Bastiaan A van den Berg[*,1,2], Jurgen Nijkamp[1,2], Marcel JT Reinders[1,2], Herman J Pel[3], Liang Wu[3], Johannes A Roubos[3], Dick de Ridder[1,2]

* b.a.vandenberg@tudelft.nl

[1] The Delft Bioinformatics Lab, Delft University of Technology, [2] Kluyver Centre for Genomics of Industrial Fermentation, [3] DSM Biotechnology Center, Delft, The Netherlands

Summary

Filamentous fungi have a high protein secretion capacity that is exploited by the fermentation industry for large-scale production of both homologous and heterologous proteins. However, many newly introduced proteins, especially heterologous, are not produced at detectable levels. Being able to predict whether or not a protein can successfully be produced is of great use for strain development.

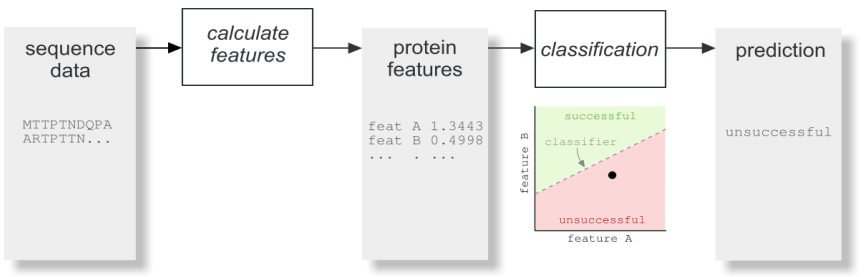


Figure 1 | **Protein classification**

Methods

We trained a classifier that predicts whether or not an *Aspergillus niger* protein will be successfully produced and secreted. The sequence data of a protein and the corresponding gene was used to calculate features. The classification algorithm used these features to predict the class to which the protein belongs, either “successful” or “unsuccessful” (Figure 1).

The classifier was trained using a set of over 600 homologous constructs that were introduced in *A. niger*, all with the same strong promoter. For each item in the dataset, the sequence data, and a binary score for successful overexpression in *A. niger* were provided.

Amongst others, we used the amino acid composition of the whole protein sequence, both for individual amino acids and for sets of amino acids (for example, a set of aromatic amino acids) as features. Features with low Mahalanobis distance between the two classes were removed from the feature set.

We used a 10-fold cross-validation loop to train and validate our classifier (Figure 2). Within each cross-validation loop, the data set is split into a training and a test set (1).

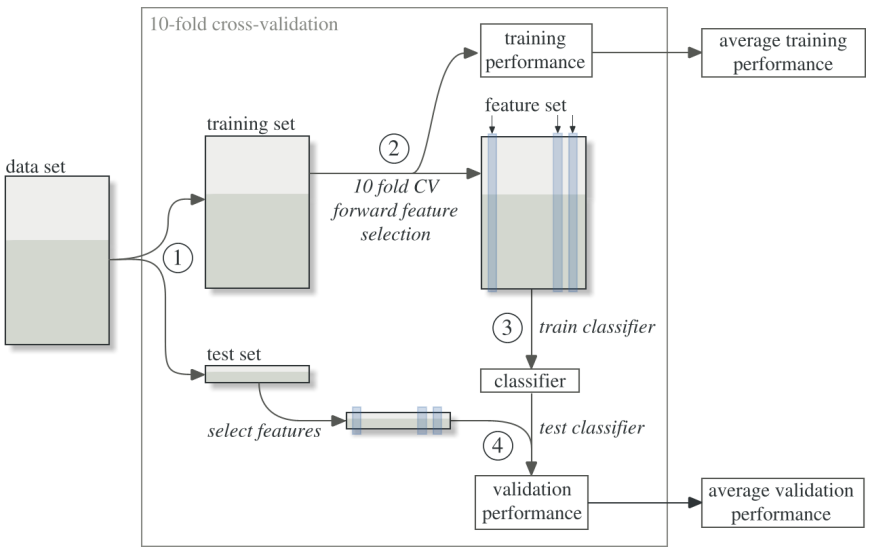


Figure 2 | **Classifier training and validation protocol**

Forward feature selection is performed on the training set (2). The selected features are used to train a classifier (3). To estimate performance, the classifier is validated on the test set, which was not employed to train the classifier (4). The 10 resulting performances are combined into an average validation performance.

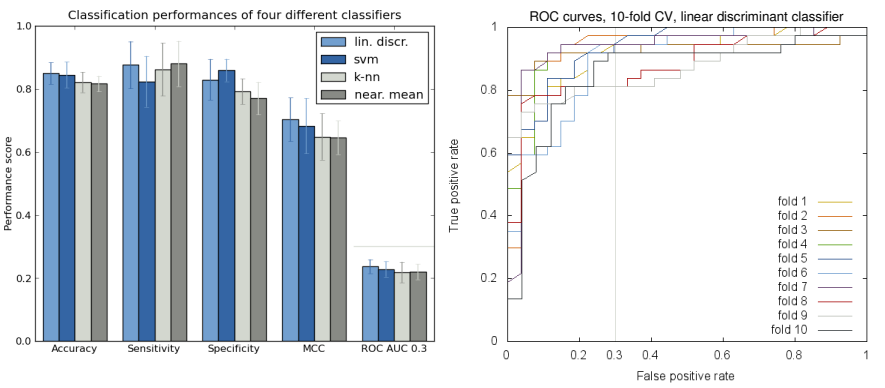
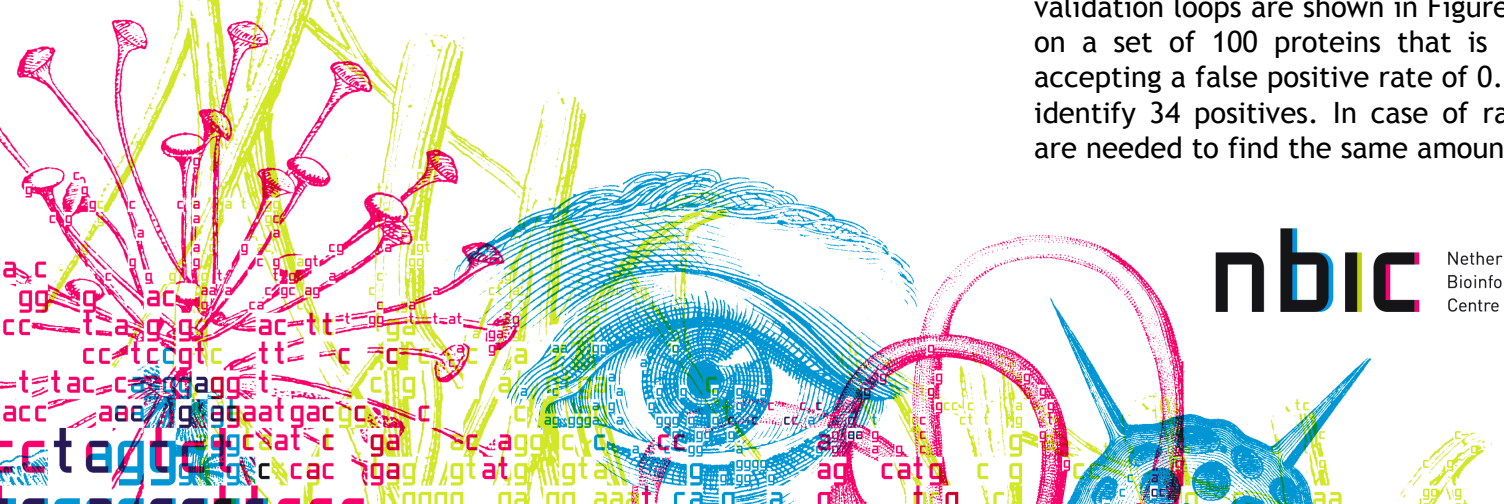


Figure 3 | ***A)* Avg. performances *B)* ROC curves best classifier**

Results

Using a linear discriminant classifier, we obtained an average accuracy of 0.85 (Figure 3A). The ROC-curves of the 10 cross-validation loops are shown in Figure 3B. When using our classifier on a set of 100 proteins that is similar to our data set, and accepting a false positive rate of 0.15, 40 lab tests are needed to identify 34 positives. In case of random selection, 80 lab tests are needed to find the same amount of positives.



nbic Netherlands Bioinformatics Centre

TU Delft

DSM