Sequence-Based Prediction of Protein Secretion Success in *Aspergillus niger*

Bastiaan A. van den Berg[1,2,4], Jurgen F. Nijkamp[1,4], Marcel J.T. Reinders[1,2,4], Liang Wu[3], Herman J. Pel[3], Johannes A. Roubos[3], and Dick de Ridder[1,2,4]

[1] The Delft Bioinformatics Lab, Delft University of Technology, The Netherlands
[2] Netherlands Bioinformatics Centre, The Netherlands
[3] DSM Biotechnology Center, The Netherlands
[4] Kluyver Centre for Genomics of Industrial Fermentation, The Netherlands
b.a.vandenberg@tudelft.nl

Abstract. The cell-factory *Aspergillus niger* is widely used for industrial enzyme production. To select potential proteins for large-scale production, we developed a sequence-based classifier that predicts if an over-expressed homologous protein will successfully be produced and secreted. A dataset of 638 proteins was used to train and validate a classifier, using a 10-fold cross-validation protocol. Using a linear discriminant classifier, an average accuracy of 0.85 was achieved. Feature selection results indicate what features are mostly defining for successful protein production, which could be an interesting lead to couple sequence characteristics to biological processes involved in protein production and secretion.

Keywords: *Aspergillus niger*, protein secretion, sequence-based prediction, classification.

1 Introduction

The filamentous fungus *Aspergillus niger* has a high secretion capacity, which makes it an ideal cell-factory widely used for industrial production of enzymes [11]. Selecting proteins for large-scale production requires testing for successful over-expression and protein secretion. Because many proteins are of potential interest, a large amount of lab work is needed. This can be reduced by developing a software tool to prioritize proteins in advance. Such a tool might also indicate which gene or protein characteristics influence successful over-expression and secretion.

Various sequence-based classifiers have been developed, for example, to predict protein crystallization propensity [6], protein solubility [8], and protein subcellular localization [14], [4]. Subcellular localization predictors have been used to predict protein secretion [16], [5], but these methods predict if a protein is inherently extracellular, whereas our aim is to predict *successful* secretion of a protein after over-expression.

In this work, we present a classifier to predict if a homologous protein will successfully be secreted after over-expression in *A. niger*, using 25 sequence-based features and providing an accuracy of 0.85.

T.M.H. Dijkstra et al. (Eds.): PRIB 2010, LNBI 6282, pp. 3–14, 2010.
© Springer-Verlag Berlin Heidelberg 2010

2 Materials and Methods

2.1 Data Set

The data set D contained 638 homologous proteins from *A. niger* CBS 513.88 [13] with a signal sequence predicted by SignalP [12]. For each protein, the open reading frame (ORF) and a binary score for successful over-expression was given. To obtain this binary success score, each protein was over-expressed through introduction of the predicted gene using the same strong glucoamylase promoter P_{GlaA}. Cultures were grown in shake-flasks and the filtered broth was put on an SDS-PAGE gel. Successful over-expression was defined as the detection of a visible band in this gel. D contained 268 successfully detected proteins (D_{pos}), and 370 unsuccessfully detected proteins (D_{neg}). The data set will be publicly available soon.

2.2 Sequence-Based Features

For each item $i \in D$, a feature vector $\boldsymbol{d_i}$ with 39 sequence-based features was constructed (Table 1). Next to simple compositional features, features that relate to protein solubility and membrane binding were chosen, because it is expected that these characteristics influence successful protein secretion. Features are calculated using the entire ORF sequence and corresponding protein sequence, including the signal peptide. A two-sample t-test with pooled variance estimation was used as class separability criterion to evaluate the performance of each feature. Features with p-value > 0.001 (gray features in Table 1) were removed, resulting in a set of 25 features used for classifier development.

For this set of features, a heat map of the hierarchical clustered (complete linkage) feature matrix is shown in Fig. 1, in which each row is a protein in D and each column is a feature. The two additional columns on the right depict the measured and predicted class labels. They show that clustering of the proteins, using this feature set, already provides a separation of D_{pos} and D_{neg}.

Compositional Features. Given a protein sequence, its amino acid composition is defined as the number of occurrences of the amino acid (frequency count) divided by the sequence length, providing 20 features. The same was done for the nucleotide composition of the coding region, providing 4 features.

Additionally, we calculated the compositions of amino acid sets that share a common property. Given a protein sequence and an amino acid set, the amino acid set composition is defined as the sum of the frequency counts of each of the specified amino acids, divided by the sequence length. Eight sets were used: helix $\{I,L,F,W,Y,V\}$, turn $\{N,G,P,S\}$, sheet $\{A,E,L,M\}$, charged $\{R,D,C,E,H,K,Y\}$, small $\{A,N,D,C,G,P,S,T,V\}$, tiny $\{A,G,S\}$, basic $\{R,K,H\}$, and acidic $\{N,D,E,Q\}$. One nucleotide set was used: GC.

As final compositional feature we used the codon adaptation index (CAI)[15], which was calculated with the codon usage index of all genes in the *A. niger* genome.

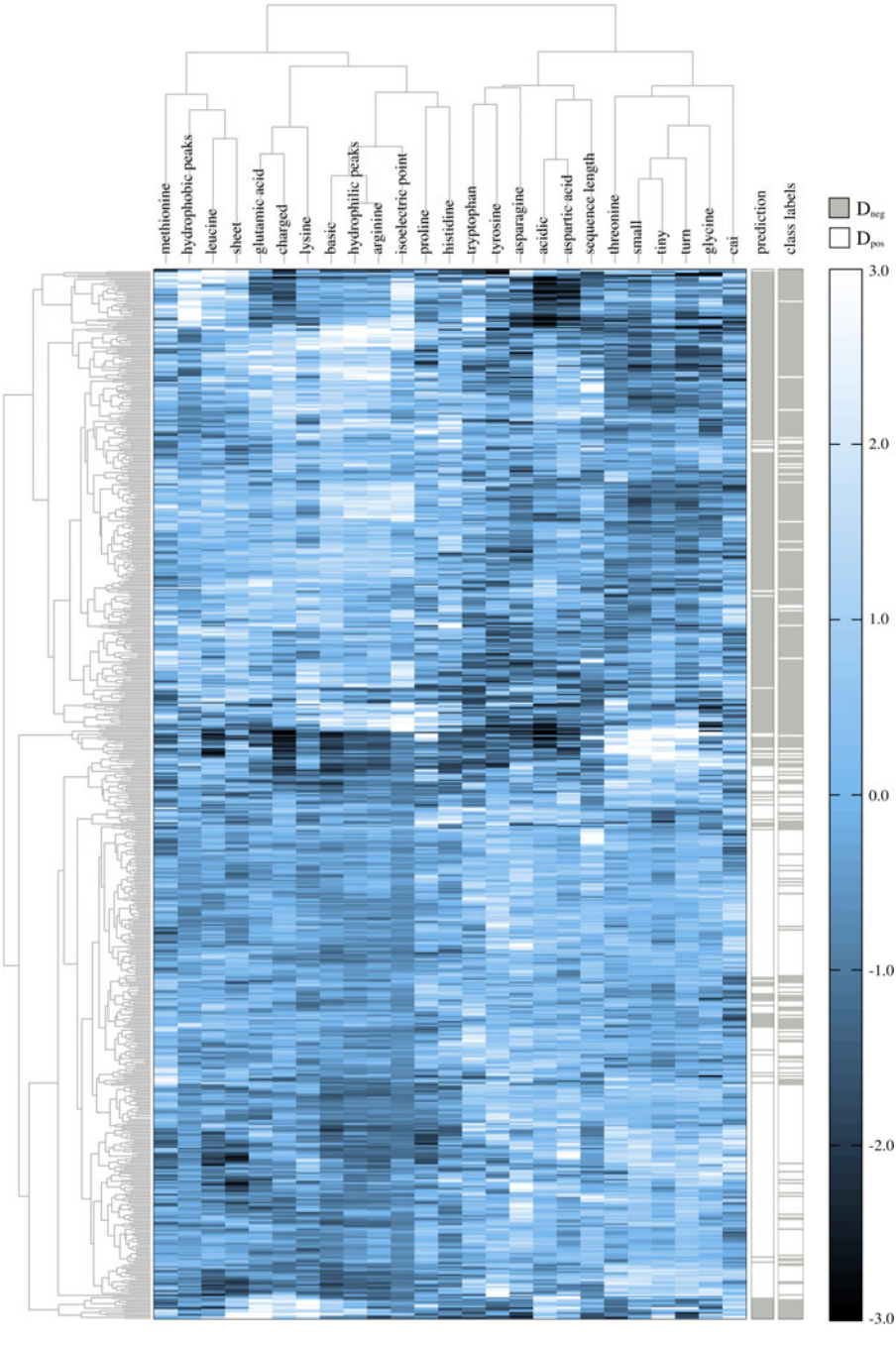


Fig. 1. Heat map of clustered feature matrix. The rows are the proteins in D and the columns are the 25 features used for classifier development. The two columns on the right depict the predicted and measured class labels respectively.

Table 1. Calculated features with class separability score

Nucleotide compositional	guanine (2.5) adenine (0.4) thymine (2.3) cytosine (2.9)	GC (1.3) CAI (5.3)	
Amino acid compositional	alanine (2.3) arginine (13.6) asparagine (15.0) aspartic acid (7.2) cysteine (0.2) glutamic acid (5.6) glutamine (0.2) glycine (9.2) histidine (4.2) isoleucine (0.9)	leucine (9.0) lysine (9.3) methionine (6.3) phenylalanine (0.1) proline (5.4) serine (1.6) threonine (8.3) tryptophan (6.3) tyrosine (13.6) valine (1.9)	helix $_{\{I,L,F,W,Y,V\}}$ (0.4) turn $_{\{N,G,P,S\}}$ (8.9) sheet $_{\{A,E,L,M\}}$ (10.8) acidic $_{\{N,D,E,Q\}}$ (7.9) basic $_{\{R,K,H\}}$ (15.7) charged $_{\{R,D,C,E,H,K,Y\}}$ (5.6) small $_{\{A,N,D,C,G,P,S,T,V\}}$ (9.7) tiny $_{\{A,G,S\}}$ (3.5)
Signal-based features	hydrophobic peaks (9.1) hydrophilic peaks (15.5)		
Global features	GRAVY (1.8) isoelectric point (16.2) sequence length (5.4)		

Signal-based Features. Two features capture the occurrence of local hydropathic peaks: *hydrophobic peaks* and *hydrophilic peaks*, both derived from a protein hydropathicity signal [1] that was constructed using the (normalized) hydropathicity amino acid scale of Kyte and Doolitle [7].

An *amino acid scale* is defined as a mapping from each amino acid to a value. Given a protein sequence, a hydropathicity signal was obtained by replacing each residue by its amino acid scale value (Fig. 2A). The signal was smoothed through convolution with a triangular function (Fig. 2B). To capture the extreme values of the smoothed signal, an upper and lower threshold were set (Fig. 2C). *Hydrophobic peaks* is defined as the sum of all areas above the upper threshold divided by the sequence length, *hydrophilic peaks* is defined as the sum of all areas below the lower threshold divided by the sequence length.

The window size and edge of the triangular function (Fig. 2B), and both thresholds (Fig. 2C) can be varied. In each CV loop of the training and validation protocol (Section 2.4), an exhaustive search was applied to optimize the features' class separability score, using: $window\ size = 3, 5, \ldots, 21$; $edge = 0.0, 0.2, \ldots, 1.0$; $threshold = 0.5, 0.54, \ldots, 0.86$ for *hydrophobic peaks* and 0.5, 0.45, ..., 0.05 for *hydrophilic peaks*.

Global Features. Three global features were used: the grand average of hydrophobicity (GRAVY), i.e., the sum of all Kyte and Doolitle amino acid scale values divided by the sequence length; the isoelectric point (pI), i.e., the predicted pH at which the net charge of the protein is zero; and finally the sequence length, i.e., the number of residues in the protein sequence.

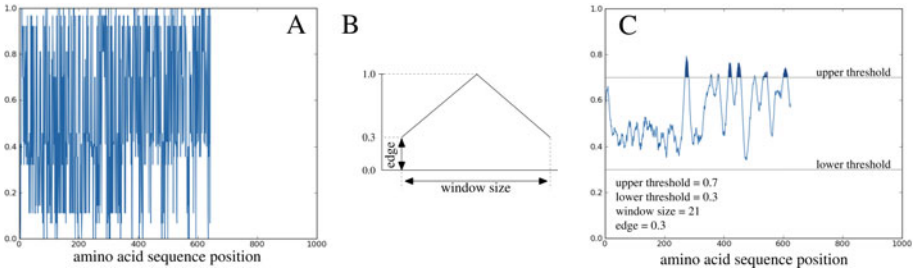


Fig. 2. Hydropathic peaks features. A) A raw protein hydropathicity signal obtained by replacing each amino acid in the sequence by its value in the normalized Kyte and Doolitle amino acid scale. **B)** Triangular function used to smooth the raw signal. **C)** Smoothed signal obtained by convolution of the raw signal in A with the function in B.

WoLF PSORT. To test whether using predicted localization would improve performance, WoLF PSORT [4] was used to predict secretion of the proteins in D. Next to the amino acid composition and the sequence length, which we also used as features, WoLF PSORT uses features based on sorting signals and functional motifs. To use the prediction as feature, we assigned proteins with intracellular localization prediction a value of 0, and proteins predicted to be extracellular a value of 1.

2.3 Performance Evaluation

We used five measures to evaluate classification performance. Four of these are based on the confusion matrix. This matrix contains the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Let the set of positives be $P = TP + FN$, the set of negatives $N = TN + FP$, the set of predicted positives $P' = TP + FP$, and the set of predicted negatives $N' = TN + FN$. The confusion matrix-based measures are; $accuracy = (TP+TN)/(P+N)$, $sensitivity = TP/P$, $specificity = TN/N$, and Matthews correlation coefficient score $MCC = (TP \times TN - FP \times FN)/\sqrt{P \times N \times P' \times N'}$. The MCC-score [9] is suited in case of different class sizes, which applies in our case. The score ranges from 0 for random assignment, to 1 for perfect prediction.

The aforementioned scores take into account only one operating point on the receiver operating characteristic (ROC) curve. As a fifth measure, we took the area under the ROC curve (AUC), thereby taking into account a range of operating points. Because the goal is to reduce the amount of lab work, we are mainly interested in low false positive rates, i.e., the left region of the ROC-curve. Therefore, we used the AUC over the range of 0 – 0.3 false positive rate (ROC0.3) as main performance measure.

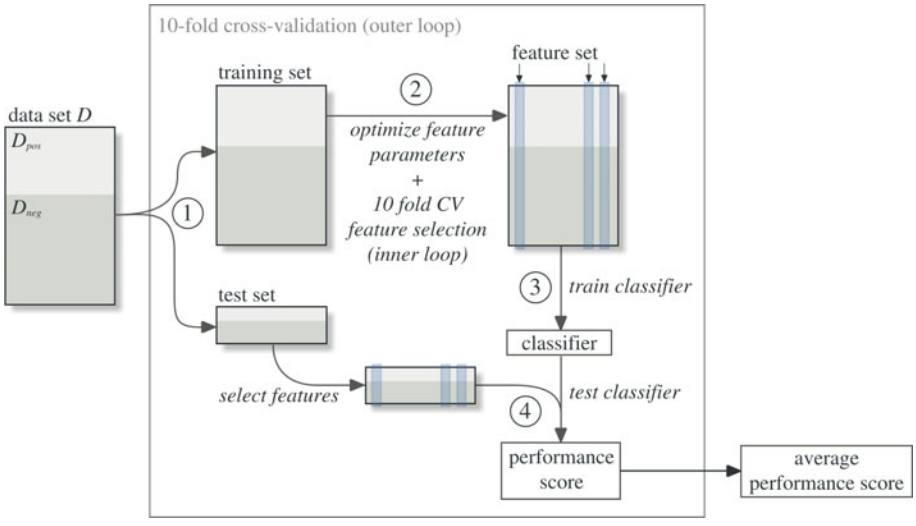


Fig. 3. Training and validation protocol

2.4 Training and Validation Protocol

To avoid overestimation of classification performance, a double 10-fold CV protocol was used, based on the protocol in [17]. We used 10-fold CV feature selection with classifier performance as selection criterion, in which the expected error $((FP/P + FN/N)/2)$ was used as performance measure.

The protocol is shown in Fig. 3. The dataset D is split into ten equal-sized random stratified sets. In each outer loop, one of the sets is used as test set, and the remaining nine as the training set (1). An exhaustive search is done to optimize the parameters of the hydropathic peaks features for maximal class separability, and 10-fold CV feature selection (inner loop) is applied on the training set to select an optimal feature set (2). As feature selection methods, we used both forward and backward feature selection. The optimal feature set is used to train a classifier on the entire training set (3). The resulting classifier is applied to the test set that was not employed for training, resulting in a performance score (4). Finally, the performance scores of the 10 CV loops are averaged, resulting in an average performance score.

The training and validation protocol was implemented in Matlab, using the PRTools pattern recognition toolbox [3].

2.5 Classifiers

We tested 8 classifiers: linear and quadratic normal density-based Bayes classifiers (ldc, qdc); nearest mean classifier (nmc); k-nearest neighbor classifier, both with $k = 1$ and with k optimized by leave-one-out CV (1nnc, knnc), naive Bayes classifier (naivebc), Fisher's least square linear classifier (fisherc), and a radial

basis support vector machine (svm, $\gamma = 1/\text{number of features}$). We used libsvm [2] for the support vector machine.

3 Results

The classifier performance scores are given in Table 2. We compared the ROC0.3 scores of the different methods using a paired t-test ($p < 0.05$) on the results of the 10 CV loops. This showed that the nearest neighbor classifiers perform significantly worse than all other methods, except for qdc with forward feature selection. The best performance was obtained with ldc and backward feature selection.

Table 2. Classifier performance scores

classifier		ROC0.3	sensitivity	specificity	MCC	accuracy
ldc	f[1]	0.232 ±0.03	0.877 ±0.08	0.819 ±0.06	0.691 ±0.08	0.843 ±0.04
	b[2]	**0.236** ±0.03	0.873 ±0.08	0.830 ±0.05	0.700 ±0.07	0.848 ±0.03
svm	f	0.228 ±0.03	0.847 ±0.08	**0.857** ±0.02	**0.701** ±0.07	**0.853** ±0.03
	b	0.232 ±0.02	0.843 ±0.08	0.854 ±0.04	0.695 ±0.09	0.850 ±0.04
fisherc	f	0.234 ±0.03	0.873 ±0.08	0.819 ±0.06	0.688 ±0.08	0.842 ±0.04
	b	0.235 ±0.02	0.881 ±0.09	0.822 ±0.05	0.698 ±0.07	0.846 ±0.03
naivebc	f	0.224 ±0.03	0.854 ±0.08	0.800 ±0.05	0.649 ±0.09	0.823 ±0.04
	b	0.230 ±0.03	0.888 ±0.08	0.803 ±0.03	0.684 ±0.07	0.839 ±0.03
qdc	f	0.221 ±0.03	0.877 ±0.06	0.803 ±0.04	0.674 ±0.06	0.834 ±0.03
	b	0.227 ±0.03	0.884 ±0.05	0.805 ±0.04	0.682 ±0.08	0.838 ±0.04
nmc	f	0.227 ±0.03	**0.910** ±0.07	0.773 ±0.04	0.678 ±0.06	0.831 ±0.02
	b	0.224 ±0.02	0.899 ±0.07	0.773 ±0.04	0.666 ±0.05	0.826 ±0.02
knnc	f	0.218 ±0.03	0.858 ±0.09	0.770 ±0.06	0.624 ±0.10	0.807 ±0.05
	b	0.214 ±0.02	0.862 ±0.06	0.778 ±0.06	0.635 ±0.05	0.813 ±0.03
1nnc	f	0.195 ±0.04	0.798 ±0.09	0.781 ±0.09	0.578 ±0.15	0.788 ±0.07
	b	0.190 ±0.03	0.809 ±0.09	0.749 ±0.08	0.557 ±0.10	0.774 ±0.05

[1] forward feature selection, [2] backward feature selection

Fig. 4 shows the ROC0.3 scores of ldcs trained on each of the 25 single features, on all 25 features, and on features obtained by backward feature selection. The classifiers are ordered by score. A paired t-test ($p < 0.001$) on the 10 CV loops showed that all single-feature classifiers are significantly outperformed by both multi-feature classifiers. Although using all features provides a higher average score than using backward feature selection, the paired t-test ($p < 0.05$) indicates that the difference is not significant.

Applying WoLF PSORT on our dataset provided a sensitivity of 0.96 and a specificity of 0.49. It appears that WoLF PSORT is too optimistic, providing a large amount of FPs. This could be explained by the difference in the problems we address; WoLF PSORT predicts extracellular proteins, whereas our method also includes successful protein production and secretion. This means that extracellular proteins in D, which are positives for WoLF PSORT, can be part

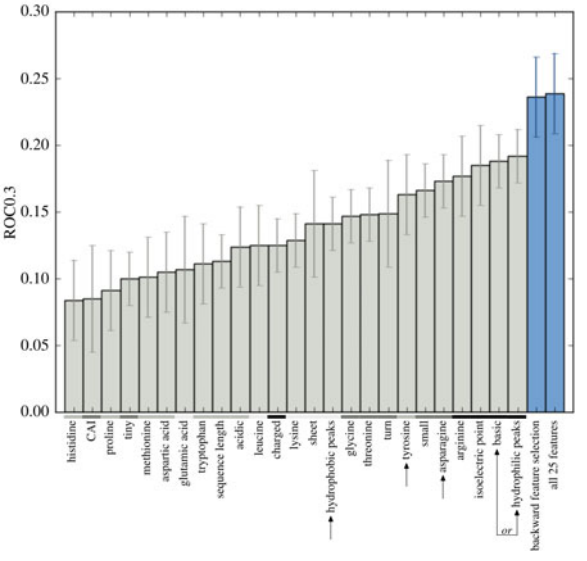


Fig. 4. Single-feature and multi-feature classification scores

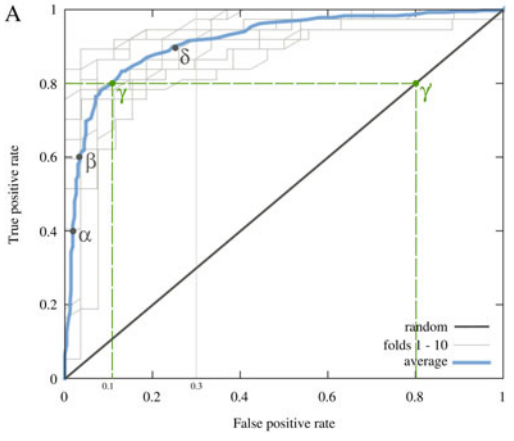
of D_{neg} because of unsuccessful protein production. We used the localization prediction as additional feature. Using ldc with backward feature selection, no significant improvement was observed, probably because the feature contains redundant data.

3.1 Operating Point Example

Fig. 5A shows the ROC of the ldc with backward feature selection. One could use this classifier to screen a set of proteins for potential over-expression candidates. For example, if we have a set S of 100 proteins that we want to screen, containing 42 positives (S_{pos}) and 58 negatives (S_{neg}) (i.e., the same fraction of positives and negatives as D), and if we use γ as operating point, a true positive rate of 0.8 will be obtained. In this case, the classifier will predict 34 true positives and 6 false positives, which means that only 40 lab experiments are needed to identify 34 positives. Without the classifier, to identify 34 positives, both the false and the true positive rate will be 0.8 (operating point γ'). In this case, 80 lab experiments will be needed to identify 34 positives, which means that the classifier could reduce the amount of lab work by a factor two (Fig. 5B).

3.2 Feature Optimization

Fig. 6 shows the optimal parameter settings for the hydrophilic and hydrophobic peaks feature as obtained in one of the CV loops. For both features, the same optimum was observed in each CV loop.



B

Example different operating points:
set of 100 proteins; 42 positives, 58 negatives

		α	β	γ	δ
	TP-rate	0.40	0.60	***0.80***	0.90
	TP	17	25	***34***	38
				γ	
Classifier	FP-rate	0.02	0.04	***0.10***	0.25
	FP	1	2	***6***	15
	lab tests	18	27	***40***	53
				γ	
Random selection	FP-rate	0.40	0.60	***0.80***	0.90
	FP	23	35	***46***	52
	lab tests	40	60	***80***	90
$\text{lab tests}_{\text{random}}$ / $\text{lab tests}_{\text{classifier}}$		2.22	2.22	***2.00***	1.70

Fig. 5. ROC-curve. A) Average ROC curve of the ten CV loops (ldc, backward feature selection). The light gray curves are the ROC curves of the separate CV loops. The diagonal line illustrates the random selection ROC curve. **B)** Numeric example that shows the amount of lab work that could be saved for different operating points.

Interestingly, when using the optimal parameter settings, the raw signal of the hydrophilic peaks is not smoothed. With $window\ size = 3$ and $edge = 0.0$, the value at a specific location in the sequence is simply the amino acid scale value of the amino acid at that specific location. Therefore, the feature is actually the same as the GRAVY feature, but using an amino acid scale in which all values greater than the threshold are set to zero, and all other values are set to the threshold minus the value. In this case, arginine is set to 0.1, lysine to 0.33, and the rest of the amino acids is set to zero. From another perspective, this feature can be seen as an amino acid set composition for the set {arginine, lysine} in which the arginine has a higher weight.

It is questionable if the resulting feature is still related to the proteins hydrophilic character. Since both arginine and lysine are also basic amino acids, it could just as well be related to the proteins basic character. Furthermore, because of the small window size, the feature does not take into account sequence order. However, it could be hypothesized that hydrophilic amino acids will mainly contribute to the proteins hydrophilic character when they have a relatively high occurrence within a larger region.

3.3 Feature Correlation

Fig. 7 shows a heat map of the hierarchical clustered (complete linkage) feature correlation matrix. The cluster at the top left shows relatively high correlations, which can be explained by the fact that the features contain redundant data: *arginine* is part of both *basic* and *charged*, *basic* is a subset of *charged*, the isoelectric point is derived from a proteins charge and therefore correlated with *charged*, and *hydrophilic peaks* takes into account the amino acids arginine and

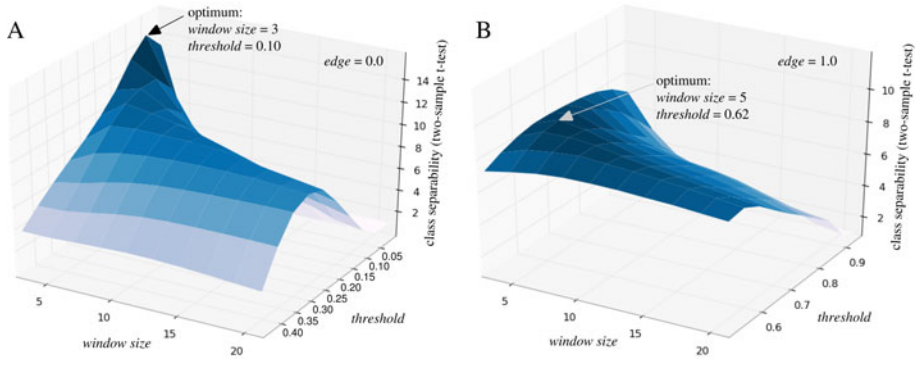


Fig. 6. Parameter optimization of hydropathic peaks features. A) Class separability scores for the hydrophilic peaks feature plotted against different parameter settings. **B)** The same as in *A*, but for the hydrophobic peaks feature. Both plots show the result for one *edge* value, different *edge* values provided similar plots. Both plots were obtained in one of the CV loops, the same optimum was found in all CV loops.

lysine, that are both in *basic* and *charged*. There is also a high correlation between *small*, *turn*, and *tiny*. This can also be explained by data redundancy: both *turn* and *tiny* are a subset of *small*.

3.4 Feature Selection

Using ldc with forward feature selection, the feature selection results of the 10 CV loops showed that: *asparagine* was always part of the top-3 selected features (7 times selected first), either *hydrophilic peaks* or *basic* was part of the top-3 selected features 9 times (6 times selected second), *hydrophobic peaks* was part of the top-4 selected features 9 times (7 times selected third), and *tyrosine* was part of the top-4 selected features 6 times (5 times selected fourth).

The high correlation between *hydrophilic peaks* and *basic* (Fig. 7), together with the fact that both have a high class separability score (Table 1), explains their mutual exclusive selection. In Fig. 4, the colors above the feature names depict what features are in the same correlation cluster and the arrows indicate what features are most often in the top-4 selected features. It shows that these features are in different correlation clusters, and are the best performing ones of their cluster. Therefore, feature selection seems to select individual features that best represent an underlying cluster of related features.

4 Discussion

To be useful for large-scale production, a protein should be produced and secreted with high yield. We report a sequence-based approach to classify proteins into *successful* or *unsuccessful* production, which was trained and validated on a set of 638 proteins. We used 10-fold CV for feature selection and classifier

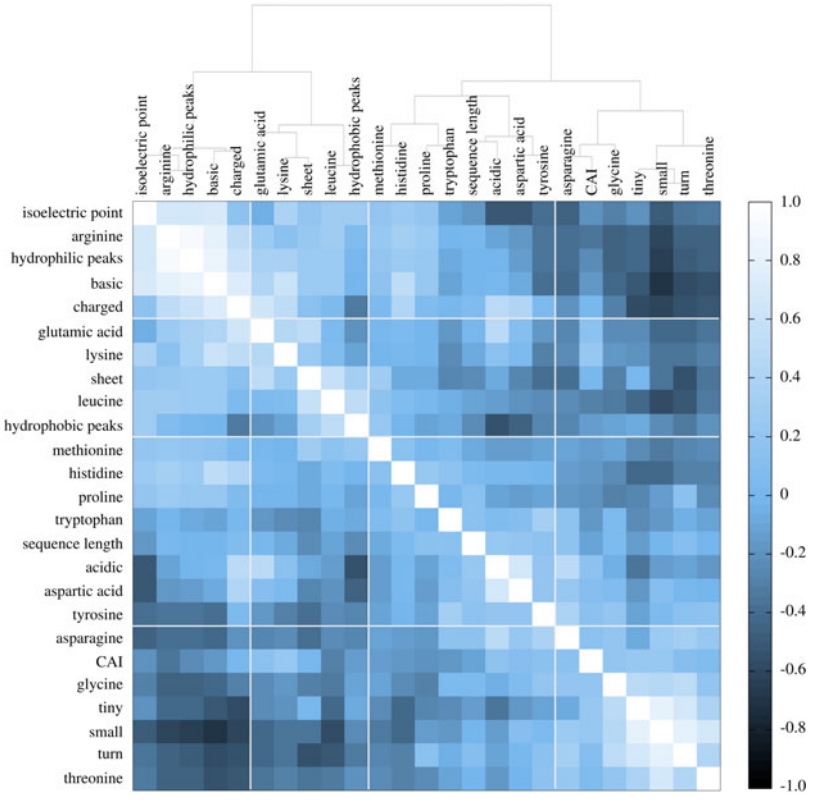


Fig. 7. Heat map of clustered feature correlation matrix

training to avoid biased performance results. Since we are mostly interested in the operating points of the first 30 percent of the ROC-curve, we used the AUC of this region as the main performance measure.

We calculated 39 features and used the 25 with highest class separability score for classification. We showed that both a classifier that uses all features and a classifier trained with feature selection, outperform classifiers trained on single features. The classifiers trained with feature selection did not significantly outperform the classifier trained on all 25 features, indicating that all features contribute to the result.

Furthermore, the feature selection results showed that asparagine, the set {arginine, lysine}, and tyrosine, as well as the hydrophobic peaks feature, were most defining in case of the linear discriminant classifier. To get more insight into protein secretion, it would be interesting to link the biological significance of these features to protein secretion mechanisms. For example, the asparagine composition could be related to N-linked glycosylation, a process that in many cases is important for protein folding and stability [10].

Acknowledgments. This work was part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI). The data set was provided by DSM Biotechnology Center.

References

1. Benita, Y., Wise, M., Lok, M., Humphery-Smith, I., Oosting, R.: Analysis of high throughput protein expression in Escherichia coli. Mol. Cell. Proteomics 5(9), 1567 (2006)
2. Chang, C., Lin, C.: LIBSVM: a library for support vector machines (2001)
3. Duin, R., Juszczak, P., Paclik, P., Pekalska, E., de Ridder, D., Tax, D., Verzakov, S.: A Matlab toolbox for pattern recognition. PRTools version 4.1, 3 (2000)
4. Horton, P., Park, K., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C., Nakai, K.: WoLF PSORT: protein localization predictor. Nucleic Acids Res. 35(Web Server issue), W585–W587 (2007)
5. Klee, E., Sosa, C.: Computational classification of classically secreted proteins. Drug Discovery Today 12(5-6), 234–240 (2007)
6. Kurgan, L., Razib, A., Aghakhani, S., Dick, S., Mizianty, M., Jahandideh, S.: CRYSTALP2: sequence-based protein crystallization propensity prediction. BMC Struct. Biol. 9, 50 (2009)
7. Kyte, J., Doolittle, R.: A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157(1), 105–132 (1982)
8. Magnan, C., Randall, A., Baldi, P.: SOLpro: accurate sequence-based prediction of protein solubility. Bioinformatics 25(17), 2200–2207 (2009)
9. Matthews, B.: Comparison of the predicted and observed secondary structure of t4 phage lysozyme. BBA-Protein Struct. 405(2), 442–451 (1975)
10. Mitra, N., Sinha, S., Ramya, T., Surolia, A.: N-linked oligosaccharides as outfitters for glycoprotein folding, form and function. Trends Biochem. Sci. 31(3), 156–163 (2006)
11. Nevalainen, K., Te'o, V., Bergquist, P.: Heterologous protein expression in filamentous fungi. Trends Biotechnol. 23(9), 468–474 (2005)
12. Nielsen, H., Engelbrecht, J., Brunak, S., Von Heijne, G.: Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng., Des. Sel. 10(1), 1 (1997)
13. Pel, H., de Winde, J., Archer, D., Dyer, P., Hofmann, G., Schaap, P., Turner, G., de Vries, R., Albang, R., Albermann, K., et al.: Genome sequencing and analysis of the versatile cell factory Aspergillus niger CBS 513.88. Nat. Biotechnol. 25(2), 221–231 (2007)
14. Pierleoni, A., Martelli, P., Fariselli, P., Casadio, R.: BaCelLo: a balanced subcellular localization predictor. Bioinformatics 22(14), e408–e416 (2006)
15. Sharp, P.M., Li, W.H.: The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15(3), 1281 (1987)
16. Tsang, A., Butler, G., Powlowski, J., Panisko, E., Baker, S.: Analytical and computational approaches to define the *Aspergillus niger* secretome. Fungal Genet. Biol. 46(1), S153 (2009)
17. Wessels, L., Reinders, M., Hart, A., Veenman, C., Dai, H., He, Y., van't Veer, L.: A protocol for building and evaluating predictors of disease state based on microarray data. Bioinformatics 21(19), 3755–3762 (2005)