SPiCE: Sequence-based Protein Classification and Exploration

Supplementary Data

Bastiaan A. van den Berg, Marcel J.T. Reinders, Johannes A. Roubos, and Dick de Ridder

1 Delft Bioinformatics Lab, Department of Intelligent Systems, Faculty Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands,
2 DSM Biotechnology Center, Delft, The Netherlands,
3 Netherlands Bioinformatics Centre, Nijmegen, The Netherlands,
4 Kluyver Centre for Genomics of Industrial Fermentation, Delft, The Netherlands

* to whom correspondence should be addressed
Aspergillus niger high-level secretion example

This example project uses the data set from our previous work\(^1\), composed of 345 *Aspergillus niger* proteins. All proteins in the data set have a predicted signal peptide (SignalP 3.0) and do not contain an ER retention signal or predicted transmembrane regions (TMHMM, phobius). The proteins were over-expressed behind a strong constitutive promoter, to test for successful high-level production and secretion. A protein was labeled successful (*pos*, 178 proteins) in case of a visible band on gel, and unsuccessful (*neg*, 167 proteins) otherwise.

The SPiCE website was used to calculate the amino acid composition, which was in turn used to: obtain *t*-statistics (Fig. S1), visualize histograms for the 4 features with the highest absolute *t*-value (Fig. S2), visualize the clustered heatmap of the full feature matrix (Fig. S3), and obtain classification performance of a linear SVM (Fig. S4).

---

Fig S2: Screenshot of the \textit{A. niger} high-level secretion example project, showing the histograms of the 4 amino acid composition features with the highest absolute $t$-values.
Fig S3: Heat map showing the feature matrix with the 245 proteins of the *A. niger* high-level secretion example project as rows and their amino acid composition feature values as columns. Both the rows and the columns are hierarchically clustered. Additionally, the protein labels are shown in the right column, with gray for unsuccessful high-level secretion and white for successful high-level secretion. The fact that the labels form clusters indicates that these features can be useful for classification.
**Fig S4**: This screenshot shows classification results for the *A. niger* high-level secretion example project. In this case, a 10-fold cross-validation classification performance of 0.837 area under the ROC-curve is achieved with a linear SVM using the amino acid composition as features.
Yeast expression example

The yeast expression example is constructed from a yeast protein expression data set. CDHit was used to reduce sequence redundancy so that none of the protein sequences share more than 35% sequence identity. The list of proteins was sorted by their observed protein molecules per cell and the top and bottom thousand proteins were taken as the high and low class.

The SPiCE website was used to calculate the amino acid composition, which were in turn used to: obtain t-statistics (Fig. S5), visualize histograms for the 5 features with the highest absolute t-value (Fig. S6), visualize the clustered heatmap of the full feature matrix (Fig. S7), and to obtain classification performance of a linear SVM (Fig. S8).

![SPiCE website](image)

**Fig S5:** Screenshot of the yeast-expression example project, showing the table with t-values for the amino acid composition features.

---


3 [http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi?cmd=cd-hit](http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi?cmd=cd-hit)
**Fig S6:** Screenshot of the yeast-expression example project, showing the histograms of the 5 amino acid composition features with the highest absolute t-values.
Fig S7: Heat map showing the hierarchically clustered feature matrix with the 2000 proteins of the yeast-expression example project as rows and their amino acid composition feature values as columns. The protein labels are in the right column, with gray for low expression and white for high expression.
Fig S8: A classification performance of 0.794 area under the ROC-curve is achieved with a linear SVM using the amino acid composition as features.