Relating amino acid patterns to successful high-level protein secretion in *Aspergillus niger*

B.A. van den Berg[*1,3,4], M. Hulsman[1,4], M.J.T Reinders[1,3,4], L. Wu[2], H.J. Pel[2], J.A. Roubos[2], D. de Ridder[1,3,4]

[1] The Delft Bioinformatics Lab, Faculty of Electrical Engineering, Mathematics & 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

* *b.a.vandenberg@tudelft.nl*

Introduction

Aspergillus niger is widely used for industrial enzyme production. Knowledge on high-level protein secretion could be useful to improve production rates. We used sequence-based classification methods to relate amino acid patterns to successful high-level secretion.

Methods & Results

Success of high-level secretion of 416 over-expressed homologous proteins was tested in the lab and used as data set. To obtain defining 'generalized' amino acid patterns, we aimed to improve similarity within a class (Fig.1).

We developed a *hierarchical feature construction* method that clusters amino acids (Fig.2). The resulting clusters seem to correspond to physicochemical clusters (Fig.3). Some interesting generalized amino acid patterns that are predictive for successful high-level secretion were found (Fig.4).

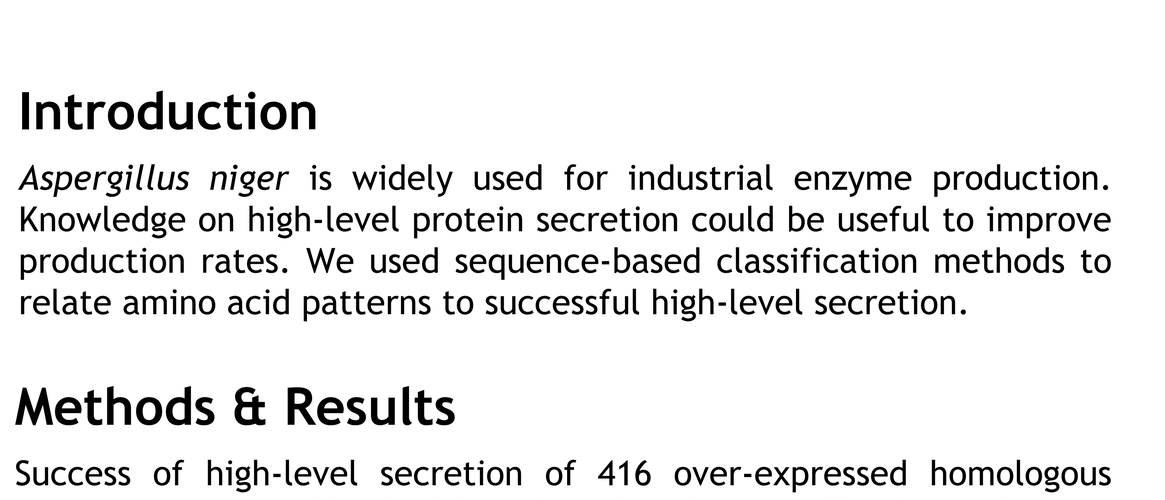


Figure 1 | **Increase sequence similarity within a class** - ***a)*** Occurrance of patterns AR (red) and AK (blue) is predictive. However, the predictor will consider sequences 1 and 2 (containing mostly AR) unequal to sequences 3 and 4 (containing mostly AK). ***b)*** Combining amino acids R and K improves sequence similarity in the positive class.

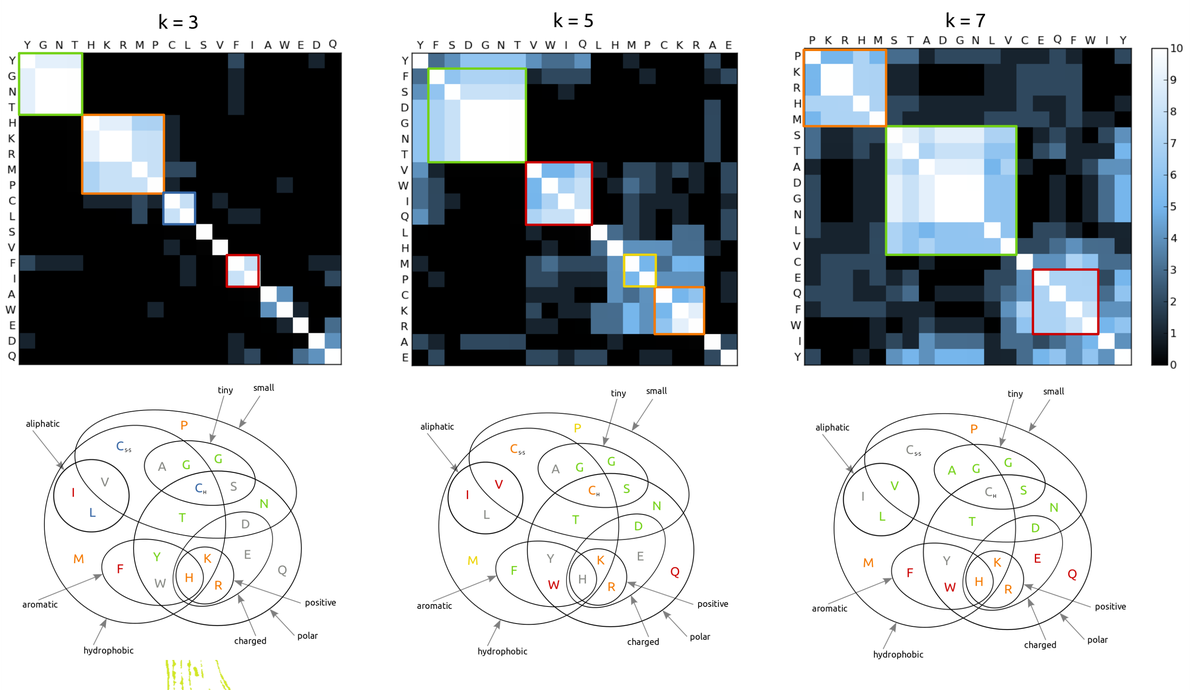


Figure 3 | **Optimal clusters for *k*-mer length 3, 5, and 7** - Number of times amino acids end up in same cluster in a 10-fold cross-validation loop (black: never, white: always). Amino acids that appeared in the same cluster 5 times or more are highlighted (coloured boxes).

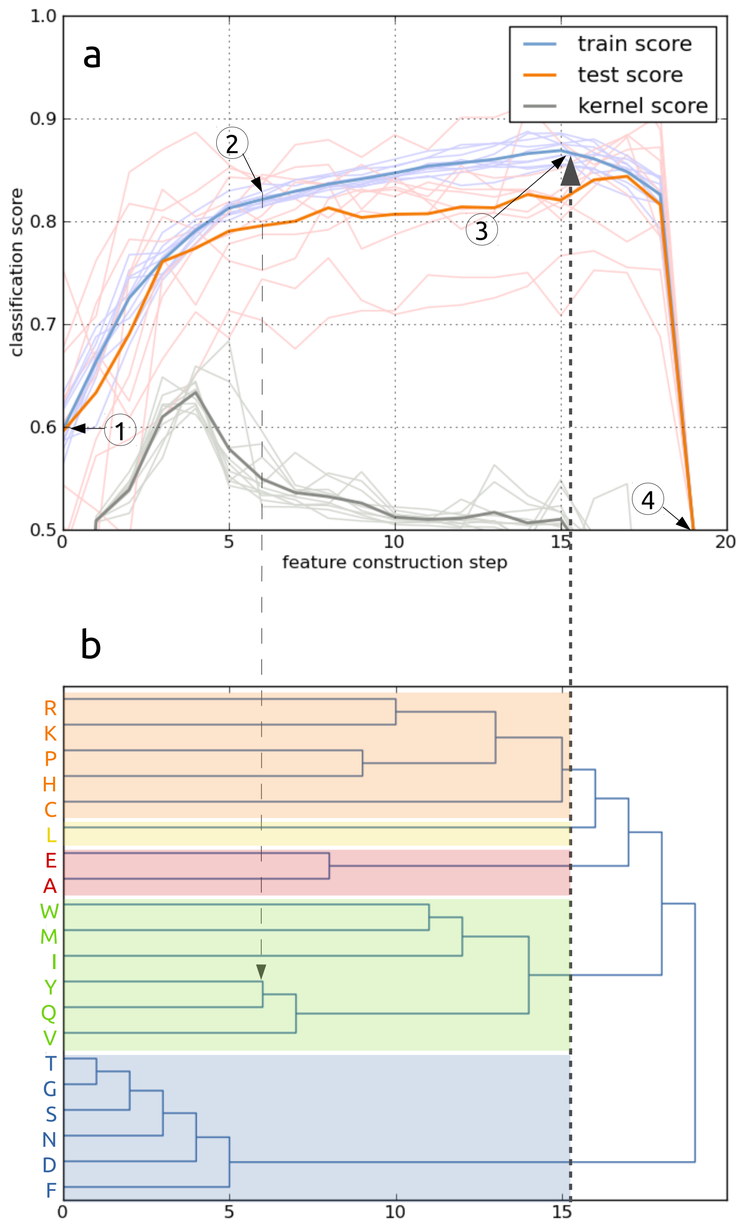
Conclusions

We showed that hierarchical feature construction can be used to obtain generalized amino acid patterns predictive for successful high-level secretion. The amino acid clusters found by the method seem to correspond to known physicochemical clusters, indicating biological relevance. Currently we are investigating if the occurence of the patterns correlate to specific structural regions, e.g. if a pattern consistently occurs in a helix region. In a later stage, this knowledge could possibly be used to rationally redesign proteins for improved secretion.

Figure 2 | **Hierarchical feature construction *k* = 5**

a) Performance is obtained using the original amino acid alphabet ***(1)***. Next, performance is obtained for each possible combined pair of letters. The best performing pair is definitively combined into one cluster. In the seventh step letters *Y* and *Q* were combined ***(2)***. This procedure is repeated until the amino acid alphabet is reduced to a single letter ***(4)***. Optimal clusters are obtained at maximal training performance ***(3)***.

b) A dendrogram showing which clusters are combined at each feature construction step (left to right). *G* and *T* are combined in the first step, *{G,T}* and *S* in the second, and so on. The colouring denotes the obtained optimal five clusters, i.e. the optimal alphabet contains five letters.



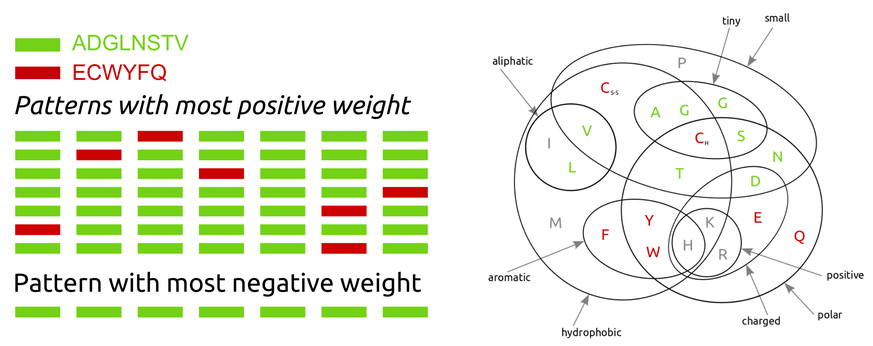
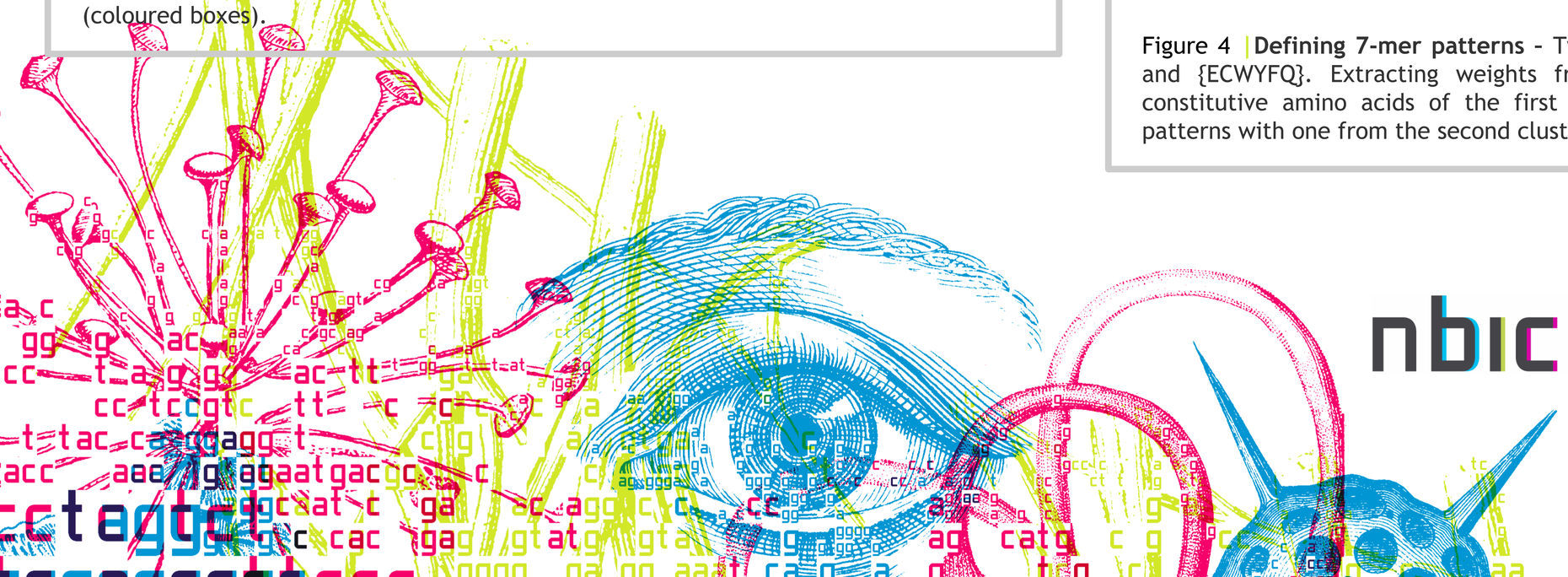


Figure 4 | **Defining 7-mer patterns** - Two obtained clusters for *k* = 7 are {ADGLNSTV} and {ECWYFQ}. Extracting weights from a trained classifier showed that the 7 constitutive amino acids of the first cluster has the most negative weight, while patterns with one from the second cluster has the most positive weights.



nbic netherlands bioinformatics centre

TU Delft

DSM BRIGHT SCIENCE. BRIGHTER LIVING.

Kluyver CENTRE | Kluyver Centre for Genomics of Industrial Fermentation