Relating sequence properties to protein secretion

B.A. van den Berg*[1,3,4], M.J.T Reinders[1,3,4], H.J. Pel[2], L. Wu[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 identify important properties for successful high-level secretion, which will be used to redesign proteins for improved secretion.

Methods & Results

Successful high-level secretion was experimentally tested for a set of homologous and a set of heterologous proteins. Support vector machines (SVMs) were trained and tested on both data sets using many different features; amino acid composition was found to be predictive.

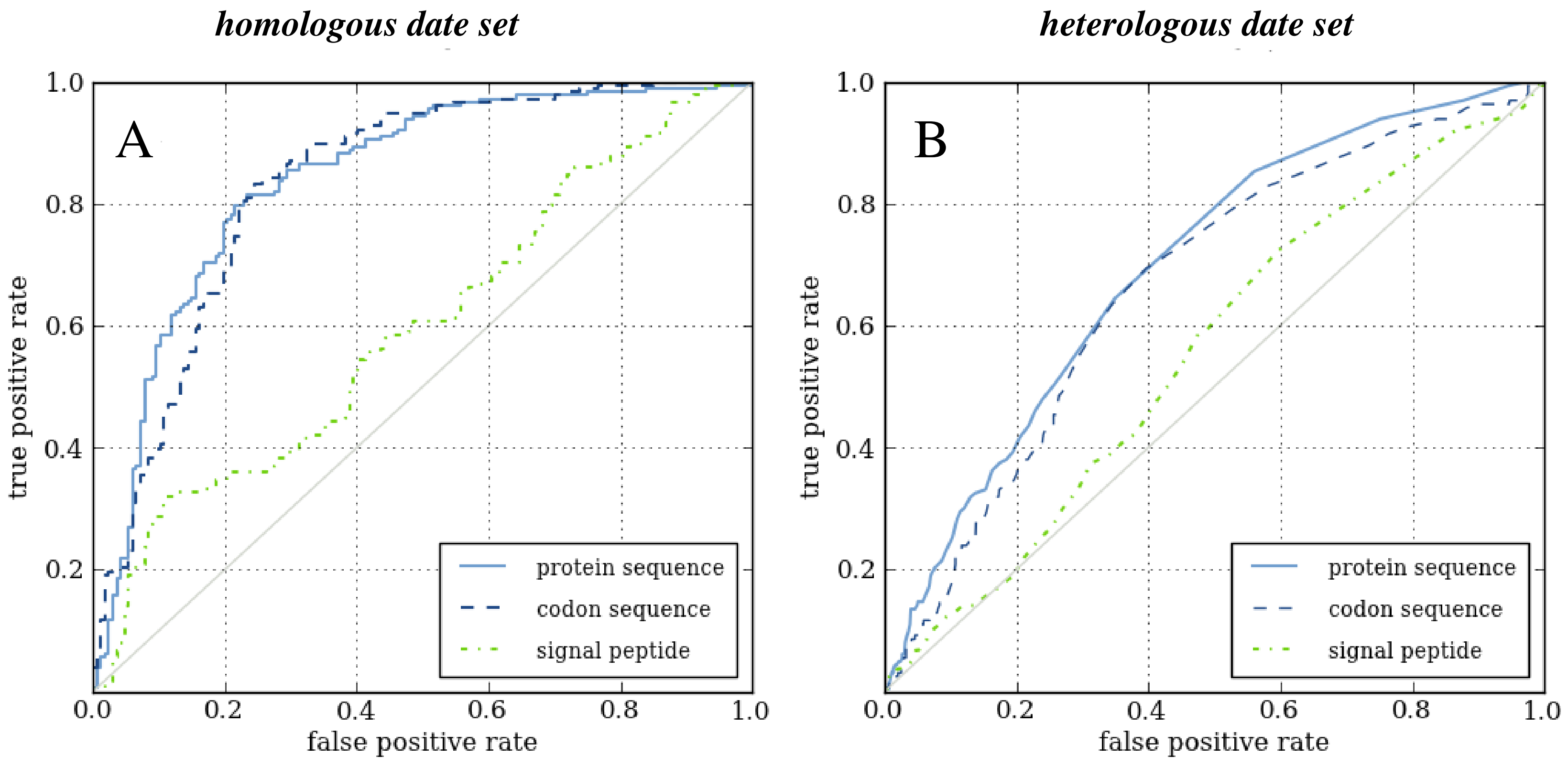

Figure 1 | **Classification performance** - ROC-curves show the classification performance on ***A)*** the homologous data set and ***B)*** the heterologous data set, that were obtained using a 10-fold cross-validation protocol. Performances are shown for SVMs that took the amino acid composition of the protein sequence, the codon composition of the protein sequence, and the amino acid composition of the N-terminal signal peptide as input.

Prediction performances are shown in Fig.1. Composition of a protein's amino acid sequence and codon sequence are highly predictive, while the amino acid composition of the signal peptide has limited predictive value. Consistently, predictive performance of homologous secretion is higher than for heterologous secretion.

An SVM implicitly assigns weights to the features it is trained on, amino acid composition in our case. For interpretation, these weights were obtained from the SVM trained on the homologous data set (x-values in Fig.2) and from the SVM trained on the heterologous data set (y-values in Fig.2). The correlation between these weights shows that both classifiers are similar.

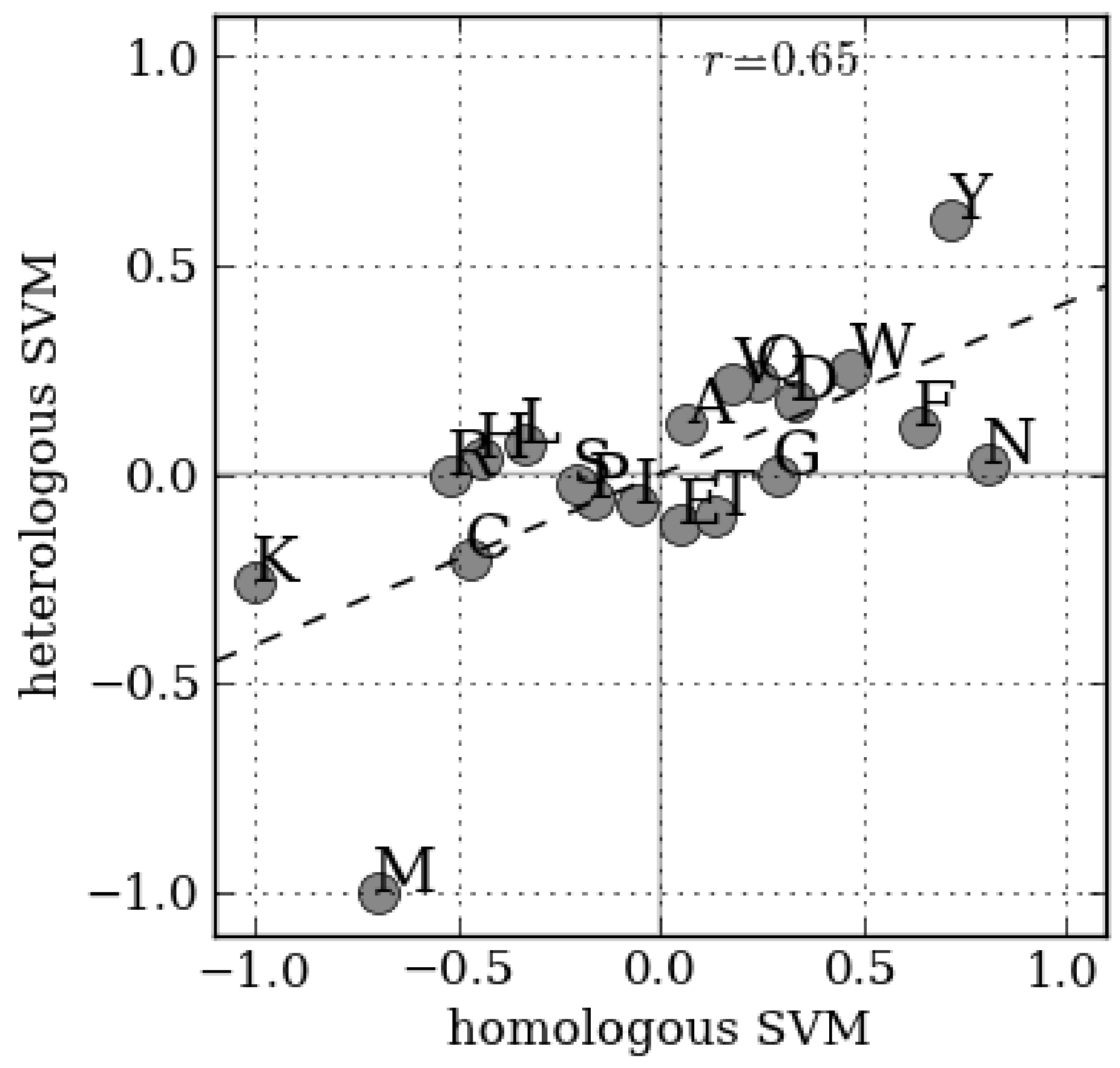

Figure 2 | **Amino acid composition weights** - The x-values and y-values are the weights obtained from the SVMs trained on protein sequences in the homologous and heterologous data set respectively.

Conclusions & Outlook

We have shown that high-level secretion of homologous protein can be accurately predicted using amino acid composition. SVMs trained on both homologous and heterologous proteins were found to be similar, showing that the same amino acids are important in both cases, i.e. general properties are defining for successful high-level secretion. However, compared to the homologous data set, a larger overlap between the positives and negatives in the heterologous data set causes a lower prediction performance.

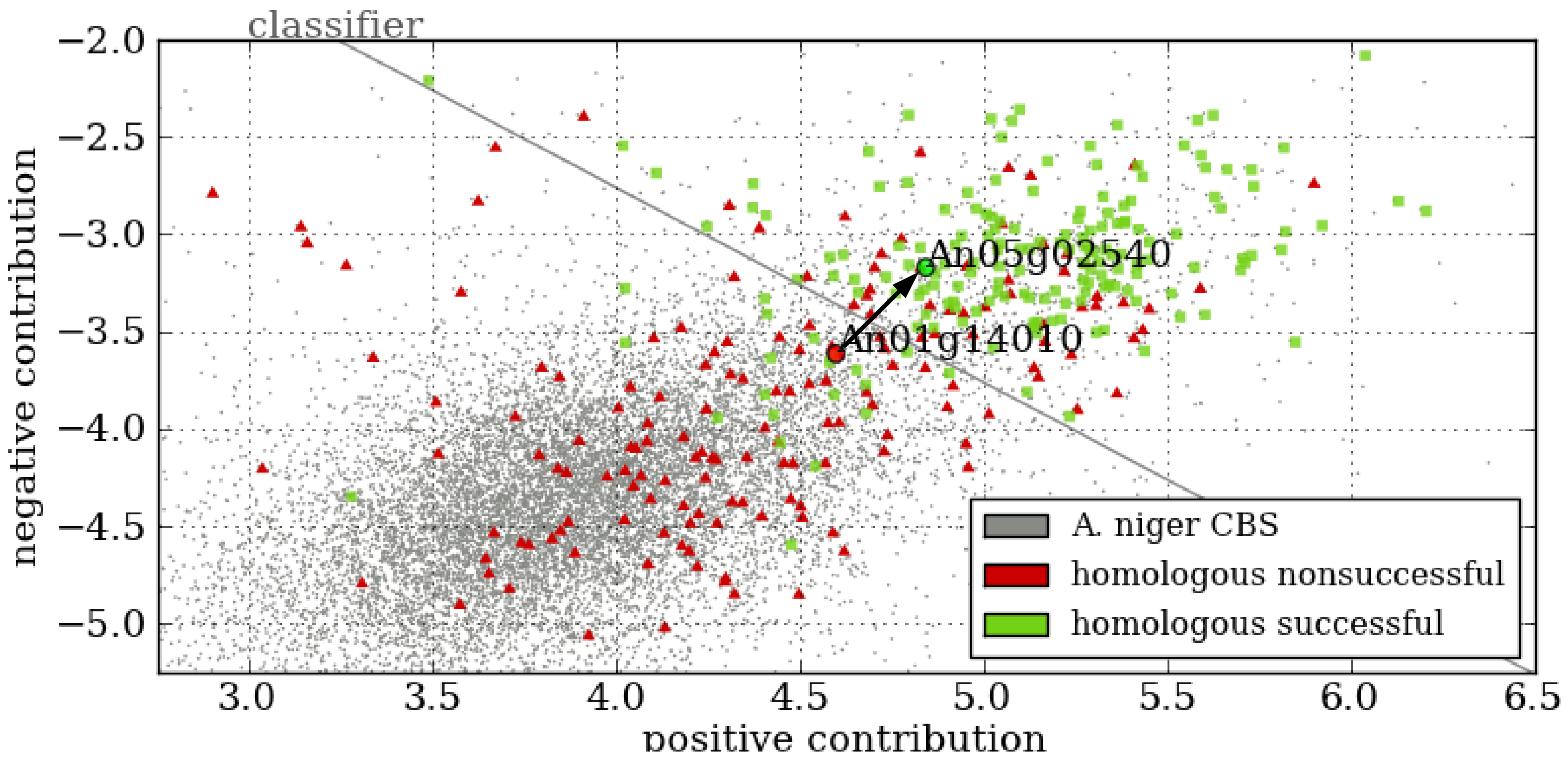

Figure 3 | **Scatter plot of the homologous data set -** For each protein, the x- and y-value are a weighted sum of its amino acid composition using the weights of the trained homologous SVM (x-values in Fig.2). The x-value is the positive contribution and y-value negative contribution. The gray line is the trained homologous SVM classifier. The labeled proteins, one successfully and one non-successfully secreted, share >50% sequence identity. This raises the question if we can improve secretion by changing a protein's amino acid composition while fixing its structure.

As a next step we aim to redesign proteins to improve secretion by adjusting the amino acid composition, thereby moving a protein to the other side of the classifier boundary, while maintaining the same protein structure (Fig.3).