OPEN ACCESS Freely available online

PLOS ONE

Exploring Sequence Characteristics Related to High-Level Production of Secreted Proteins in *Aspergillus niger*

Bastiaan A. van den Berg[1,2,4]*, Marcel J. T. Reinders[1,2,4], Marc Hulsman[1,2,4], Liang Wu[3], Herman J. Pel[3], Johannes A. Roubos[3], Dick de Ridder[1,2,4]

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

Abstract

Protein sequence features are explored in relation to the production of over-expressed extracellular proteins by fungi. Knowledge on features influencing protein production and secretion could be employed to improve enzyme production levels in industrial bioprocesses via protein engineering. A large set, over 600 homologous and nearly 2,000 heterologous fungal genes, were overexpressed in *Aspergillus niger* using a standardized expression cassette and scored for high versus no production. Subsequently, sequence-based machine learning techniques were applied for identifying relevant DNA and protein sequence features. The amino-acid composition of the protein sequence was found to be most predictive and interpretation revealed that, for both homologous and heterologous gene expression, the same features are important: tyrosine and asparagine composition was found to have a positive correlation with high-level production, whereas for unsuccessful production, contributions were found for methionine and lysine composition. The predictor is available online at http://bioinformatics.tudelft.nl/hipsec. Subsequent work aims at validating these findings by protein engineering as a method for increasing expression levels per gene copy.

Citation: van den Berg BA, Reinders MJT, Hulsman M, Wu L, Pel HJ, et al. (2012) Exploring Sequence Characteristics Related to High-Level Production of Secreted Proteins in *Aspergillus niger*. PLoS ONE 7(10): e45869. doi:10.1371/journal.pone.0045869

Editor: Gustavo Henrique Goldman, Universidade de Sao Paulo, Brazil

Received April 21, 2012; **Accepted** August 22, 2012; **Published** October 1, 2012

Funding: This work was supported by the BioRange programme of the Netherlands Bioinformatics Centre (NBIC) and was part of the Kluyver Centre for Genomics of Industrial Fermentation, subsidiaries of the Netherlands Genomics Initiative (NGI). This funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The used data set was provided by DSM Biotechnology Center. Study design, data collection and analysis, decision to publish, and preparation of the manuscript has been done in cooperation with the DSM employees who are co-authors.

Competing Interests: DSM is a Life Sciences and Material Sciences company with enzyme products in its portfolio. Three of the co-authors, JAR, HP, and LW, are DSM employees. IP on this or related work might be filed. The used data set was provided by DSM Biotechnology Center. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: b.a.vandenberg@tudelft.nl

Introduction

In industrial enzyme production, high-level protein production and secretion are key requirements. The commercial market value was estimated to be nearly US$ 5 billion in 2009; roughly half of production is accounted for by filamentous fungi [1]. Interest in industrial enzymes is still growing, driven by the increased demand for sustainable production processes and the need to move from a fossil fuel-based to a bio-based economy. This calls for the exploration of novel enzymes, as well as predictable methods for high-yield production processes. The filamentous fungi *Aspergillus niger*, *Aspergillus oryzae* and *Hypocrea jecorina* are the major fungal workhorses in industrial enzyme production, due to their efficiency in producing polysaccharide-degrading enzymes (particularly amylases, pectinases, lipases and xylanases) in high amounts. The genome sequence of the enzyme producing *A. niger* strain CBS513.88 was published in 2007 [2] and compared with a related citric-acid producing strain ATCC1015 in 2011 [3].

Although rational genetic engineering strategies have been developed [4–6], including codon optimization, strong promoters etc., protein overexpression is still often an art. Heterologous expression in particular is less successful, often hampered by low production levels [7]. Although protein overexpression, including the secretion process and quality control mechanisms such as UPR-ERAD, has been studied widely [8–12], no generic solution to improve heterologous overexpression is yet available. More successful is the use of fusion proteins, at the cost of reduced overall yield due to the production of the fusion partner. We propose another strategy: to re-engineer proteins to better match the cell's production and secretion machinery. In this paper, we take a first step in this direction.

Our aim is to identify protein characteristics that correlate with the production level of secreted proteins in a library of *A. niger* strains. Ideally, data on protein structure, folding and even post-translational modification and processing, both intracellular and extracellular, should be exploited to enhance our understanding of the cellular processing of successful and unsuccessful candidates. Such data is however limited and expensive to obtain, unattainable for large sets of non-commercial proteins. On the other hand, some of this information is also captured in the protein sequence as such, which therefore should be informative. Using a large and

diverse library of protein sequences should allow focus on generic aspects, ignoring protein-specific aspects.

We constructed a unique library of over 2,600 strains to overexpress a selected protein sequence. After transformation using overexpression cassettes, productivity of each strain was screened by shake-flask growth and analysis of the protein composition of the supernatant on gel. Protein production was scored positive when, compared to the mother strain, an additional band on SDS-PAGE gel was observed in the expected molecular weight range; otherwise it was scored negative. Characteristics found to distinguish between proteins in the positive and negative classes may point to sequence features that could be adapted in optimization schemes to further "streamline" proteins that already show good expression, in analogy to what has been achieved with codon optimization, where gene sequences are adapted to match the translational machinery [13].

Statistically significant associations between sequence features and positive and negative class membership can be obtained relatively easy. However, such analyses are typically univariate, considering only individual features. In contrast, machine learning algorithms can combine large numbers of features and by that achieve more optimal prediction performance. Recently, different machine learning techniques have been applied on sequence data to predict protein localization [14–16] or protein solubility [17]. A disadvantage of machine learning approaches is that they often result in "black boxes", not easily providing insight into the properties that are defining for the prediction. With few exceptions [18,19], sequence-based predictors are rarely interpreted.

We developed a sequence-based predictor for extracellular protein production by *A. niger*, with the explicit goal of interpreting which combinations of features are most predictive. We consider a large number of potentially interesting features and develop predictors for both homologous and heterologous gene expression. Sequence data was found to be predictive for both, although less accurate prediction results were obtained for the heterologous data set. Interestingly, interpretation of the underlying model parameters show that for both data sets similar properties are predictive for extracellular protein production. The trained classifier algorithms are made available in a freely accessible online tool (http://bioinformatics.tudelft.nl/hipsec).

Methods

Experimental Setup

Proteins were experimentally tested for high-level production in *A. niger*. Binary success scores were obtained by SDS-PAGE of (at least) triplicate shake-flask samples with strains over-expressing the introduced gene as described below. A positive success score was given when a clear visible band was present, negative otherwise.

Strain. The strain used in this work is a recombinant strain derived from DS03043, a progenitor of CBS 513.88, in which the *gla*A loci (i.e., the promoter and coding sequences) were deleted, creating the so-called Δ*gla*A loci. From this strain, a strain was derived with a strongly reduced production of abundantly secreted proteins by inactivation of the major protease *pep*A and a number of alpha-amylases [20]. This protease- and amylase-reduced strain was used as host strain for over-expression of proteins.

Molecular biology techniques. In order to obtain targeted integration and expression of any desired gene in the above-mentioned host strain, a standard expression unit was used, where the gene of interest was inserted between the host-own glucoamylase promoter (original 2 kb 5′ *gla*A sequence) and glucoamylase terminator elements (original 2 kb 3′ *gla*A sequence) in a proprietary *Escherichia coli* vector. The expression unit, a linear piece of DNA, was targeted via single-crossover to the *Δgla*A locus using the homology in the 2 kb 3′- and direct downstream 2 kb 3″-*gla*A regions with the identical 2kb-left and 2kb-right flanks of the expression cassette, as described in [20]. All gene sequences were cloned in the *E. coli* vector exactly from start ATG until stop codon.

Shake flask fermentations. *A. niger* strain spores were pre-cultured in 20 ml CSL pre-culture medium (100 ml flask, baffle). After growth for $18-24$ hours at 34°C and 170 rpm, 10 ml of this culture was transferred to Fermentation Medium (FM). Fermentation in FM was performed in 500 ml flasks with baffle with 100 ml fermentation broth at 34°C and 170 rpm for the number of days indicated. The CSL medium consisted of (in amount per liter): 100 g Corn Steep Solids (Roquette), 1 g $NaH_2PO_4 \cdot H_2O$, 0.5 g $MgSO_4 \cdot 7H_2O$, 10 g glucose$\cdot H_2O$ and 0.25 g Basildon (antifoam). The ingredients were dissolved in demi-water and the pH was adjusted to pH 5.8 with NaOH or H_2SO_4; 100 ml flasks with baffle and foam ball were filled with 20 ml fermentation medium and sterilized for 20 min. at 120°C. The fermentation medium (FM) consisted of (in amount per liter): 150 g maltose$\cdot H_2O$, 60 g Soytone (peptone), 1 g $NaH_2PO_4 \cdot H_2O$, 15 g $MgSO_4 \cdot 7H_2O$, 0.08 g Tween 80, 0.02 g Basildon (antifoam), 20 g MES, 1 g L-arginine. The ingredients were dissolved in demi-water and the pH was adjusted to pH 6.2 with NaOH or H_2SO_4; 500 ml flasks with baffle and foam ball were filled with 100 ml fermentation medium and sterilized for 20 min. at 120°C.

SDS-PAGE electrophoresis. Sample pre-treatment: 30 μl sample was added to 35 μl water and 25 μl NuPAGETM LDS sample buffer ($4\times$, Invitrogen) and 10 μl NuPAGETM Sample Reducing agent ($10\times$, Invitrogen). Samples were heated for ten minutes at 70°C in a thermo mixer. SDS-PAGE was performed in duplicate according to the supplier's instructions (Invitrogen: 4–12% Bis-Tris gel, MES SDS running buffer, 35 min. runtime). One of the two gels was used for blotting, 10 μl of the sample solutions and 1 μl marker M12 (Invitrogen) were applied on the gels (NuPAGETM BisTris, Invitrogen). The gels were run at 200 V, using the XCELL Surelock, with 600 ml 20 times diluted MES-SDS buffer in the outer buffer chamber and 200 ml 20 times diluted MES-SDS buffer, containing 0.5 ml of antioxidant (NuPAGETM Invitrogen) in the inner buffer chamber. After running, the gels were fixed for one hour with 50% methanol/7% acetic acid (50 ml), rinsed twice with demineralised water and stained with Sypro Ruby (50 ml, Invitrogen) overnight. Images were made using the Typhoon 9200 (610 BP 30, Green (532 nm), PMT 600 V, 100 micron) after washing the gel for ten minutes with demineralised water. Typical detection limit for the fermentation samples using the described method is around 50 mg/l.

Data

Two protein data sets were tested for high-level production, one for homologous gene expression (Supplementary Table S1) and one for heterologous gene expression. Proteins in the heterologous data set originated from 14 different fungal donor organisms (Supplementary Table S2–S3). All proteins have a signal peptide (length >10 amino acids) as predicted by SignalP 3.0 [21], and a total sequence length longer than 100 amino acids. Proteins containing the most common ER retention signal (C-terminal [HK]DEL) and proteins predicted to be transmembrane by both TMHMM [22] and Phobius [23] were filtered out of the data set.

To avoid biasing subsequent analyses, sequence redundancy was reduced using BLASTCLUST [24]. Two sequences were considered redundant when the aligned sequences shared $>40\%$ identity over a length of minimal 90% for at least one of the

sequences. From the obtained protein clusters, we selected a representative protein, with the shortest average distance to all other proteins in the cluster, and removed the remainder. If a cluster contained proteins with both positive and negative labels, one positive and one negative protein was selected. This resulted in data sets *hom* and *het* containing 345 proteins (178 positives, 167 negatives) and 991 proteins (163 positives, 828 negatives), respectively.

To train a classifier on *hom* en test it on *het*, a data set het_{hom} was constructed that contains the *het* data set without proteins that share $>40\%$ identity with any protein in *hom*. This data set contained 906 (128 positives, 778 negatives) proteins.

Protein Representations

Figure 1 shows the ten different sequences that were used to represent a protein: r_0) the ORF codon sequence, using a 64 letter codon alphabet; r_1) the N-terminal signal peptide sequence; r_2) the mature protein sequence (excluding the signal peptide); r_3) the predicted solvent accessibility sequence, using B for buried and E for exposed; r_4) the parts of the mature protein sequence predicted to be buried, and r_5) to be exposed, both using the 20 letter amino acid alphabet; r_6) the predicted secondary structure sequence, using H for α-helix, E for β-strand, and C for random coil; r_7) the parts of the mature protein sequence predicted to be in a helix structure; r_8) in a strand structure; and r_9) in a random coil region, all three using the 20 letter amino acid alphabet.

We used randomized versions of the different structural sequences: r_4') randomized buried sequence, r_5') randomized exposed sequence, r_7') randomized helix sequence, r_8') randomized strand sequence, and r_9') randomized coil sequence, to test whether their actual amino acid content or just their length is predictive. For example, if for a given protein 50 residues are predicted to be in a helix structure, i.e. the helix sequence has length 50, a randomized helix sequence is constructed by selecting 50 residues from the entire protein sequence at random.

Structural Predictions

SignalP 3.0 [21] was used to predict N-terminal signal peptide presence and signal peptide cleavage site. From the neural network output, we used the default D-value threshold (0.43) to decide if a protein contains a signal peptide and used the predicted signal peptide cleavage site to split a protein sequence into a signal peptide part and a mature protein sequence part (Figure 1A). NetSurfP 1.0 [25] was used to predict structural location (either buried or exposed) of each amino acid in a mature protein sequence (Figure 1B). PsiPred 3.21 [26] was used to predict secondary structure of the mature protein sequence, using UniRef90 as a database (Figure 1C).

Classification

A linear support vector machine (LIBSVM [27]) was used for classification [28], in which the prediction y is a weighted

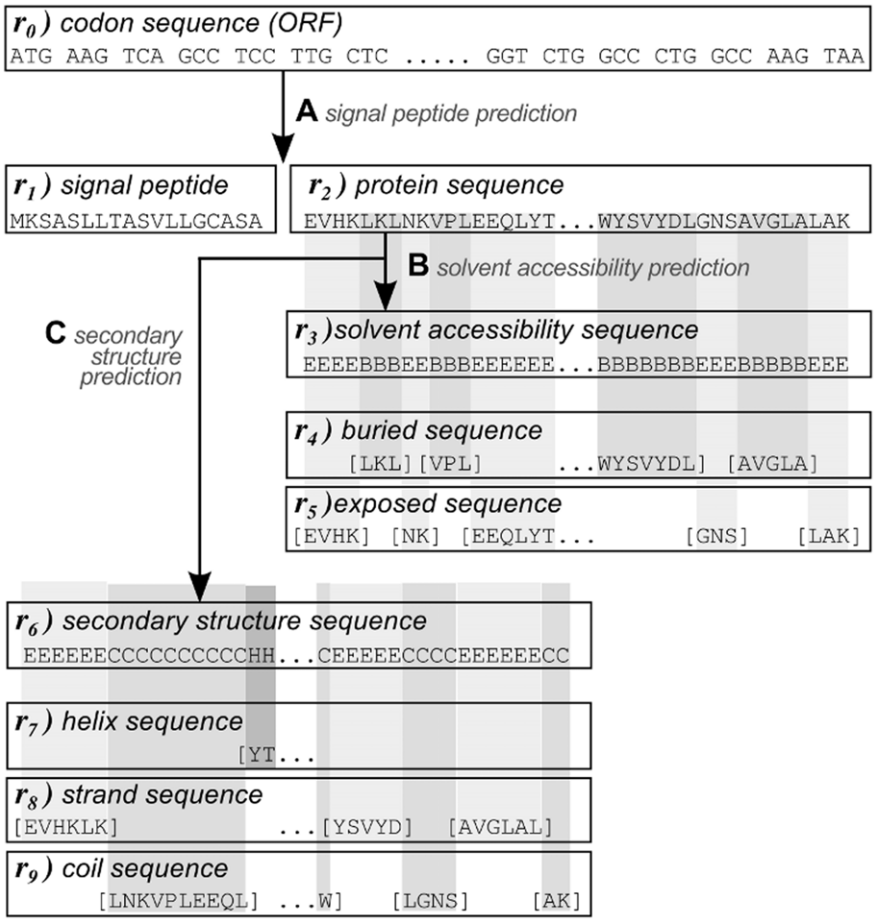

Figure 1. Different sequence-based protein representations. The different shades of gray denote predicted buried (B) and exposed (E) regions in case of the the solvent accessibility, and predicted helix (H), strand (E), and random coil (C) region in case of the secondary structure.
doi:10.1371/journal.pone.0045869.g001

combination of kernels $K(s_i,z)$ between the training objects i and a test object z:

$$y = \sum_{s_i \in S} \alpha_i y_i \Phi(s_i)\Phi(s_z) = \sum_{s_i \in S} \alpha_i y_i K(s_i, s_z) \quad (1)$$

For each object (protein) i, α_i is the weight assigned to the object as obtained from the trained classifier ($0 < \alpha_i \leq 1$ if the object is a support vector, $\alpha_i = 0$ otherwise), y_i the class label (-1 or 1), s_i the sequence of protein i, and $\Phi(s_i)$ a mapping from sequence to feature space. The SVM is trained by optimizing a quadratic programming problem:

$$\max_{\alpha} \sum_{s_i \in S} \alpha_i - \frac{1}{2} \sum_{s_i \in S} \sum_{s_j \in S} y_i y_j \alpha_i \alpha_j K(s_i,s_j) \quad \text{s.t.} \quad 0 \leq \alpha_i \leq C \; \forall i \quad \text{and} \quad \sum_{s_i \in S} \alpha_i y_i = 0 \quad (2)$$

The parameter C, controlling the trade-off between training error and classifier complexity, was optimized using a simple grid search over $1.0 \times 10^{-6}, 1.0 \times 10^{-5}, \ldots, 1.0 \times 10^{6}$. Classifier performance on a data set was estimated by running a double 10-fold cross-validation (CV) loop, in which C was optimized in an inner CV-loop on the training set. As performance measure we used the area under the receiver-operator characteristic curve (AUROC) [29]. Classifier performance is defined as the average AUROC over the CV-loops. When separate training and test sets are used, a classifier was trained on the first data set, optimizing C in a 10-fold CV-loop, and tested on the second data set, again using the AUROC as performance measure.

In the cross-validation error estimation procedure, a predictor is repeatedly trained on 90% of the data set and tested on the remaining 10% of the data set. If features derived from a training set that are important for discriminating between the positive and negative class also yield good performance on the test set, then these features apparently allow good generalization. In this sense, a good CV performance can be interpreted as an *in silico* validation of the features found.

Classifier Interpretation and Comparison

For a given set of sequences S, the feature weight vector $\mathbf{w}$ from a trained SVM classifier was obtained using:

$$\mathbf{w} = \sum_{s_i \in S} \alpha_i y_i \mathbf{\Phi}(s_i). \quad (3)$$

Classifiers were compared by taking the correlation between $\mathbf{w}$ of both trained classifiers. A high correlation indicates a high similarity between the classifiers, both assigning similar weights to the same features.

Feature Sets

We derived distinct sets of sequence-based features, f_0–f_{22}, which will be described below. A visualization of feature matrices f_0, f_1, f_2, and f_{12} for *hom* and *het* are given in Supplementary Figures S1–S8. Features f_1–f_{14} were used in an inner product kernel ($K(x, y) = x^T y$); for features f_{15}–f_{22} we used a spectrum kernel (see below).

Composition-based features. (f_0 – f_9) The composition of sequences r_0 – r_9 (Figure 1). For a sequence s on alphabet A, the composition c is defined as:

$$\mathbf{c}(s) = \frac{\text{count}(l,s)}{|s|} \quad \forall l \in A, \quad (4)$$

in which count (l,s) is a function that counts the number of occurrences of letter l in sequence s, and $|s|$ is the length of the sequence. The size of the feature vector c depends on the size of alphabet A, e.g. the composition of the codon sequence r_0 results in a feature vector of length 64 and the composition of the protein sequence r_2 results in a feature vector of length 20. This means that f_0 and f_2 consist of 64 and 20 features respectively. f_4', f_5', f_7', f_8' and f_9' are the compositional features of the randomized sequences r_4', r_5', r_7', r_8' and r_9'

f_{10}) Predefined amino acid cluster composition of r_2 using the 11 predefined clusters in Table 1. The clusters are based on those defined in [30]. (Note: In this clustering it sometimes occurs that an amino acid is both inside and outside a cluster, based on its state; e.g. a free cysteine is in the polar cluster, while a cysteine that forms a disulfide bridge is outside the polar cluster. Without structural data, amino acid states are unknown. We therefore removed an amino acid from the cluster if it also resides somewhere outside that cluster, i.e. cysteine is not considered to be part of the polar cluster.) For a sequence s and clusters G, the cluster composition vector $\mathbf{cc}$ is defined as:

$$\mathbf{cc}(s,G) = \frac{\sum_{l \in g} \text{count}(l,s)}{|s|} \quad \forall g \in G. \quad (5)$$

f_{11}) Optimized amino acid cluster composition of the protein sequence (r_2) using clusters that are optimized for our data set using the method described in the next section (Amino acid clustering).

Sequence-derived features. f_{12}) Using r_0, codon usage was calculated for the 59 codons that non-uniquely encode for an amino acid. Codon usage is defined as the codon count divided by the amino acid count of the amino acid it encodes for.

f_{13}) Four other sequence-derived features: the signal peptide length, the protein sequence length, the codon adaptation index

Table 1. Predefined amino acid clusters.

cluster	amino acids
small	V, C, A, G, T, P, S, D, N
polar uncharged	S, W, N, Q, T, Y
aromatic	F, Y, W, H
acidic	D, E
charged	H, K, R, E, D
basic	K, R, H
hydrophobic	I, L, V, M, F, Y, W, H, C, A, T, K
tiny	A, G, S
nonpolar	A, V, L, I, M, G, F, P
aliphatic	I, L, V
polar	Y, W, H, K, R, D, E, T, S, N, Q

doi:10.1371/journal.pone.0045869.t001

[31] that was calculated using a codon usage index derived from all *A. niger* genes, and the isoelectric point. The last two values were calculated using the codon sequence (r_0) and the protein sequence (r_2) respectively, both using the Biopython software package [32].

Selected features. f_{14}) A two-sample *t*-test (python SciPy package [33]), was applied to a set of 124 features, combining the features from feature sets $f_0, f_1, f_2, f_3, f_6, f_{10}$, and f_{13}. Features with a *p*-value $<1.0\times10^{-4}$ were selected for forward feature selection, 36 and 33 features for *hom* and *het* respectively (Supplementary Table S4).

In a 10-fold cross-validation loop, forward feature selection was applied on the training set. Features were added one by one, based on their prediction performance as determined using a second inner 10-fold CV-loop, until prediction performance starts to drop. To reduce calculation time, parameter *C* was not optimized but based on observations fixed to 1.0×10^{3} and 1.0×10^{-6} for *hom* and *het*, respectively. The selected features per CV-loop for both *hom* and *het* are given in Supplementary Table S5.

Pattern-based features. $f_{15}, f_{16}, \ldots, f_{22}$) We employed spectrum kernels [34], which define similarities between sequences based on fixed-length subsequence (*k*-mer) counts, as implemented by Shogun [35] to search for predictive patterns. We calculated $k=2,3,4,5$ spectrum kernels using r_2 ($f_{15}, \ldots, f_{18}$) and r_1 ($f_{19}, \ldots, f_{22}$).

Amino Acid Clustering

We developed a method that forms amino acid clusters using our data sets, thereby constructing new features optimized for our data. A cluster is defined as a set of one or more amino acids. For the resulting clusters, each amino acid can be in one cluster only, not every amino acid needs to be in a cluster.

The method starts with selecting the best performing amino acid, i.e. the amino acid that, when used as the only feature, provides the best classification performance, the same as in forward feature selection. For example, if the fraction of lysine in a protein provides the best separation between the positive and negative class, this amino acid will start the first cluster. In the next iteration, for the remaining 19 amino acids, classification performance is tested for two cases: 1) with the amino acid added as new cluster and 2) with the amino acid added to the existing cluster. In case of the example, when adding alanine, classification performance is tested both using the fraction of lysine and the fraction of alanine as two separate features, and using the sum of the fractions of lysine and alanine as a single feature. The case that provides the best classification performance is selected. In the next iteration, with 18 amino acids remaining, the same procedure is applied. This iteration cycle is proceeded until there are no more amino acids left. Finally, the overall best performing clusters are the output of the method. Consequently, it might happen that some amino acids will not be selected at all.

This procedure is implemented in a 10-fold CV-loop, obtaining the best performing clusters on the training set and using them as cluster composition features on the test set. The selection protocol is applied in an inner CV-loop to avoid biases towards the training data. The obtained clusters per CV-loop are given in Supplementary Table S6.

Statistical Pattern Discovery

The statistical motif finding approach MEME [36] was used to find patterns (described as position-dependent letter-probability matrix) that occur once in every sequence (oops mode) of a data set. Discriminative motif discovery was performed using the successful secreted proteins as input with the unsuccessful secreted proteins as negative sequences and vice versa. This was done for both *hom* and *het*. The minimal and maximal motif lengths were set to 2 and 15 respectively.

Results

Sequence Data is Predictive for High-level Protein Production

To test if the sequence data is informative, we used it to predict successful high-level protein production. Classifiers were built using an extensive set of sequence-based features. Performance results (AUROC) of 10-fold CV experiments on both *hom* and *het* are shown in Table 2, 0.5 indicating random prediction and 1.0 perfect prediction. Best classification performances of 0.85 and 0.75 AUROC respectively (boldface in Table 2) show that sequence data is predictive. As additional support, classifier outcome for the *A. niger* proteome (Supplementary Figure S11) shows an expected result, predicting successful high-level production for only a fraction of the proteome.

Considering the composition-based features, similar results were observed for the codon sequence (f_0) and the protein sequence (f_2), which is expected because of the relation between the two sequences. For *hom*, high performance using protein sequences is in line with results of our previous work [37]. Similarly, results of other previous work, regarding only protein localization and not production rate, reported different amino acid compositions for

Table 2. Prediction performance scores (AUROC).

features	*hom*→*hom*	*het*→*het*	
Composition-based features			
f_0	**0.85**	0.70	*ORF codon composition*
f_1	0.66	0.51	*signal peptide AA composition*
f_2	0.83	0.70	*mature protein AA composition*
f_3	0.68	0.51	*buried-exposed composition*
f_4 (f_4')	0.80 (0.80)	0.65 (0.64)	*buried AA composition*
f_5 (f_5')	0.82 (0.78)	0.64 (0.65)	*exposed AA composition*
f_6	0.62	0.57	*helix-strand-coil composition*
f_7 (f_7')	0.68 (0.70)	0.60 (0.57)	*helix AA composition*
f_8 (f_8')	0.70 (0.72)	0.61 (0.57)	*strand AA composition*
f_9 (f_9')	0.80 (0.80)	0.65 (0.65)	*coil AA composition*
f_{10}	0.80	0.63	*AA clusters composition*
f_{11}	0.83	0.67	*optimized AA clusters comp.*
Sequence-derived features			
f_{12}	0.64	0.54	*codon usage*
Selected features			
f_{14}	0.84	**0.75**	*feature selection*
Pattern-based features			
f_{15}	0.82	0.63	*2-mer counts protein*
f_{16}	0.77	0.61	*3-mer counts protein*
f_{17}	0.68	0.60	*4-mer counts protein*
f_{18}	0.57	0.47	*5-mer counts protein*
f_{19}	0.63	0.54	*2-mer counts signal peptide*
f_{20}	0.59	0.52	*3-mer counts signal peptide*
f_{21}	0.54	0.51	*4-mer counts signal peptide*
f_{22}	0.56	0.50	*5-mer counts signal peptide*

doi:10.1371/journal.pone.0045869.t002

intra- and extracellular proteins [38,39]. Although the codon sequence shows a slightly higher score for *hom*, it does not significantly outperform the protein sequence ($p = 0.14$ for a paired *t*-test on the test scores of the 10 CV-loops).

The predictive power of the amino acid composition of the signal peptide (f_1) proves to be limited, clearly outperformed by both the codon and protein sequence. More advanced methods, taking into account letter/pattern location [40,41], did not improve prediction results (results not shown).

Similar Characteristics are Important for Both Data Sets

Figure 2 shows the ROC-curves of the composition-based classifiers discussed thus far. Figure 2A and Figure 2B show the average result of a 10-fold CV-loop on *hom* and *het* respectively. Figure 2C shows the result of a classifier trained on *hom* and tested on *het*. Remarkably, this shows similar results as the classifiers trained on *het*, suggesting that the homologous classifier generalizes well to predict high-level production for *het*. In fact, the classifiers trained on *het* performed even slightly worse. This might be due to the fact that this data set is too heterogeneous, originating from 14 different species, which makes it harder to build a generic classifier and may have caused over-fitting in the CV-loops.

The good generalization of the *hom* classifier on the *het* data set suggests that classifiers trained on *hom* and *het* are similar, i.e. perform their predictions based on the same sequence characteristics. The correlation of 0.65 in Figure 3 shows that this is indeed the case. The figure shows the contribution of each amino acid as obtained from the *hom* and *het* classifier, both trained using the protein amino acid composition (f_2). Positive values indicate contribution to successful high-level production and negative values indicate contribution to unsuccessful high-level production.

For both *hom* and *het*, a remarkable positive and negative contribution of respectively tyrosine (Y) and methionine (M) is apparent. For *hom*, also asparagine (N) and lysine (K) show an outstanding positive and negative contribution respectively. Considering amino acid properties, it is observed that the basic and the sulfur-containing amino acids have a negative contribution whereas the (uncharged) aromatic amino acids have a positive contribution.

Besides comparing the amino acid contributions of the *hom* and *het* classifier, we also compared them to amino acid synthesis costs as defined in [42]. With the exception of the aromatic amino acids, a negative correlation is shown between the *hom* contributions and the amino acid costs (Supplementary Figures S9–S10), suggesting a preference for "cheap" amino acids for high-level secretion.

Basic and Aromatic Amino Acids are Predictive

From a structural and functional perspective, it is often more useful to look at the physicochemical properties of an amino acid, rather than looking at the 20 amino acids as different entities. Therefore, based on physicochemical properties [30], we defined 11 predefined amino acid clusters (Table 1), and used these as features (f_{10}). In this case, a correlation of 0.71 was observed between the *hom* and *het* classifier (Figure 3B). The aromatic amino acids have a high contribution to high-level production, which, looking back at Figure 3A, is similar to the amino acid contributions, except for the positively charged histidine (H). A negative contribution is observed for the basic amino acids, also consistent with the observations in Figure 3A.

Since it is unclear which amino acid clusters are suitable for what problem, we developed a novel method that uses the data set to construct clusters. The best performing clusters (f_{11}) obtained in ten CV-loops are jointly shown as a heat map in Figure 4. The non-diagonal values show the number of times that two amino acids were found in the same cluster. The diagonal values show how often an amino acid was found in any cluster.

The diagonal values correspond to the results observed in Figure 3A: highly contributing amino acids were often found in a cluster. For *het*, noteworthy exceptions are phenylalanine (F) and glycine (G), both of which always ended up in a cluster despite their low contribution.

The non-diagonal values also match the results in Figure 3A. As can be observed, amino acids with a positive contribution (green letters) and amino acids with a negative contribution (red letters) often form clusters, whereas amino acids with contradicting contributions rarely do. The occurrence of only few light cells show that not many amino acids consistently form the same cluster. Only clusters with phenylalanine (F), glycine (G), aspartic acid (D), and glutamine (Q) occur relatively often in both data sets, but those do not share an obvious physicochemical property. Despite the high contributions observed for the aromatic amino acids in Figure 3B, clustering of these amino acids occurred only a few times.

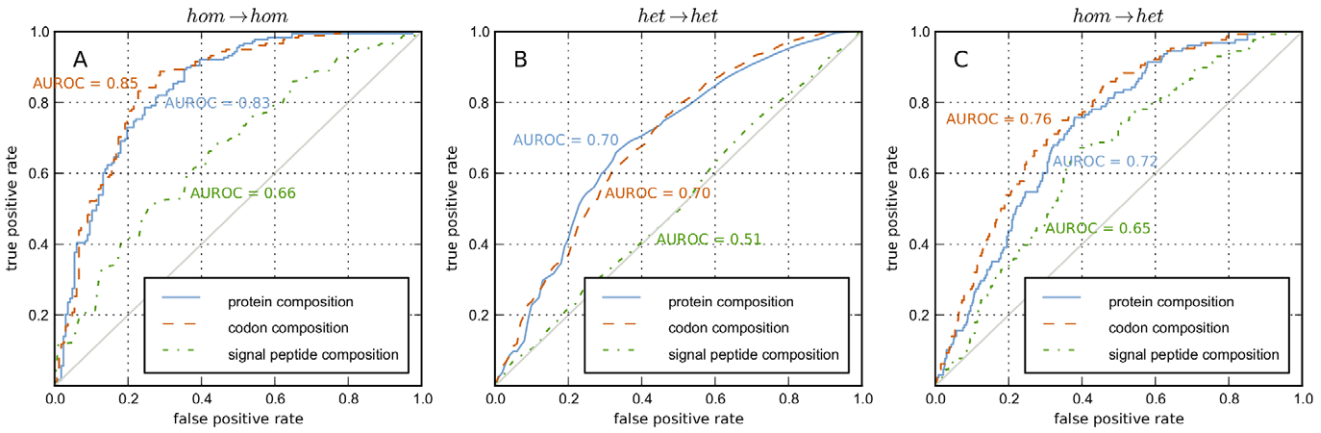

Figure 2. Classification performances. ROC-curves of composition-based classifiers using the codon sequence (f_0), the signal peptide sequence (f_1), and the protein sequence (f_2). Performances are shown for classifiers **A)** trained and tested on *hom*, **B)** trained and tested on *het*, and **C)** trained on *hom* and tested on *het*.
doi:10.1371/journal.pone.0045869.g002

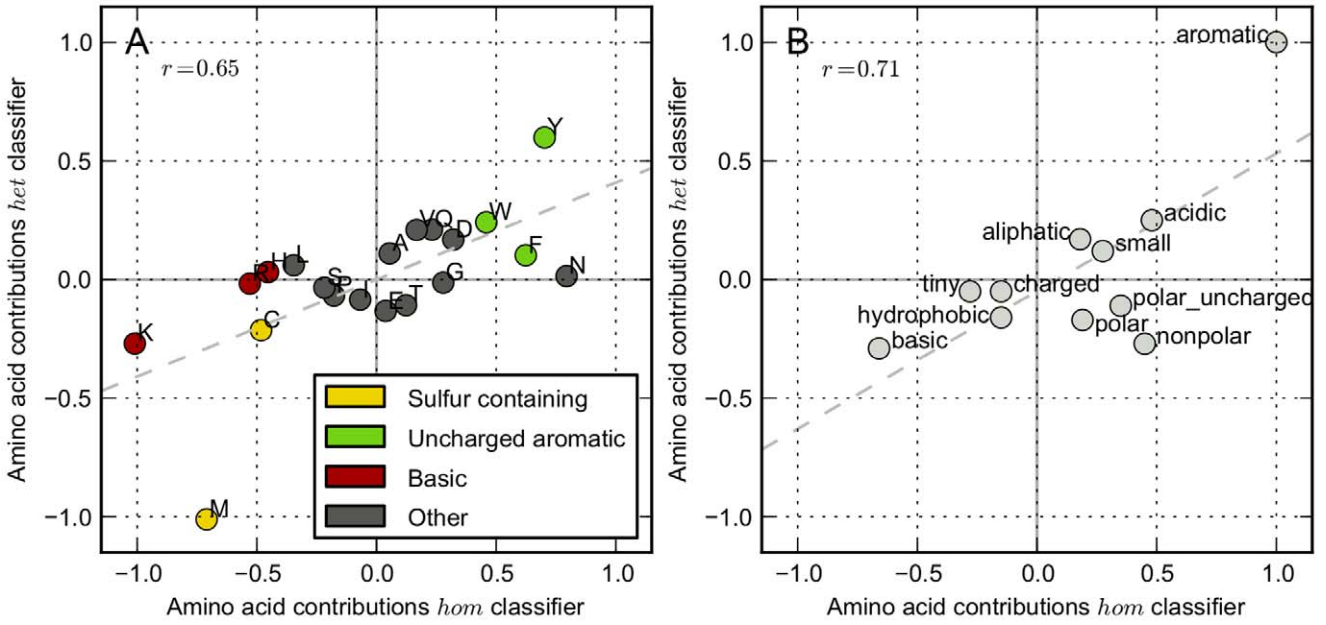

Figure 3. Comparing *hom* and *het* classifiers. Amino acid contributions obtained from *hom* and *het* trained classifiers are the x- and y-values respectively, the correlation is denoted by r. Contributions are normalized per classifier (axis): each contribution is divided by the maximum absolute contribution. The plots show the contributions obtained from classifiers trained using **A)** the protein amino acid composition (f_2) and **B)** the predefined amino acid cluster composition (f_{10}).
doi:10.1371/journal.pone.0045869.g003

Structural Subsequences have Limited Information

The secondary structure composition (f_6) shows to be little predictive, with an AUROC of 0.62 and 0.57 for *hom* and *het* respectively. Results using the amino acid composition of the helix (f_7), strand (f_8), and coil sequence (f_9) suggest that the coil sequence is more informative than the helix and strand sequence, however, a similar result was obtained using a randomized version of the sequence (f_9', score between brackets in Table 2). This indicates that the coil sequence, although it provides higher classification performance, is not more informative than the helix and strand

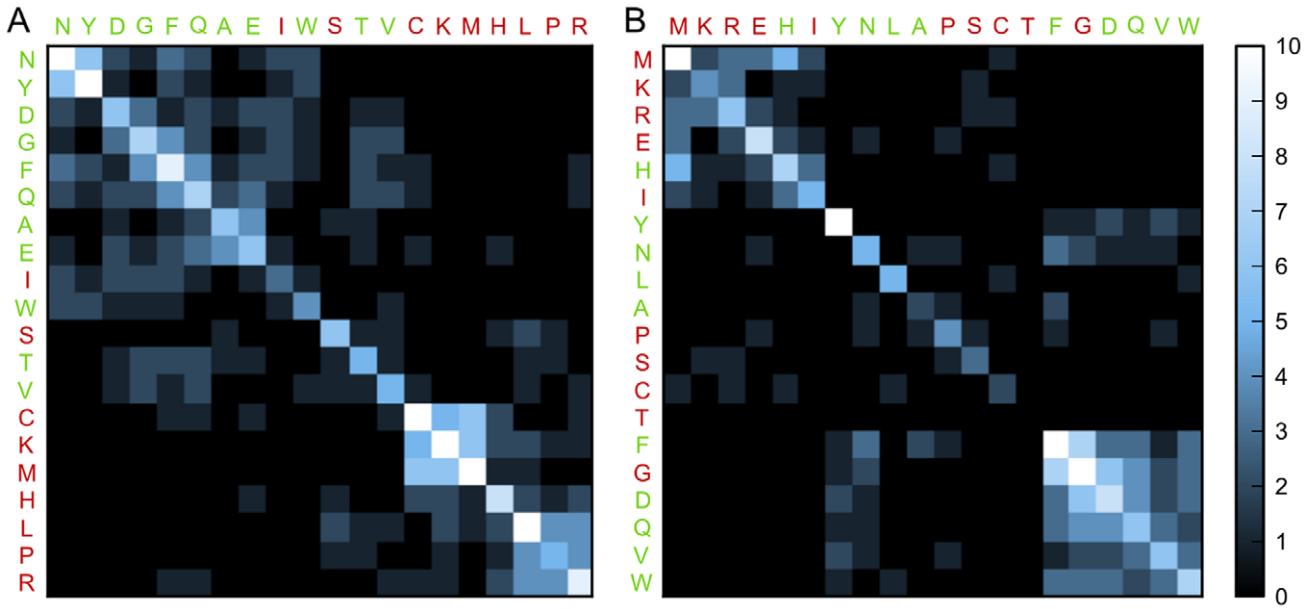

Figure 4. Best performing amino acid clusters. The heat maps show the combined result of the best performing clusters obtained in 10 CV-loops for both *hom* (**A**) and *het* (**B**). The values on the diagonals denote how often an amino acid ended up in a cluster (due to selecting the optimal clusters, amino acids might not be selected at all). The colors on the non-diagonal places denote how often two amino acids ended up in the same cluster. Complete linkage hierarchical clustering was used to cluster the heat map, using the euclidean distance as distance measure. The color of the amino acid letters indicates if the amino acid has a positive (green) or negative (red) contribution in Figure 3A.
doi:10.1371/journal.pone.0045869.g004

sequence. The better performance can be explained by the length of the sequence, proteins are on average composed of 60% coil, 20% helix, and 20% strand.

Considering the solvent accessibility, the distribution of buried and exposed amino acids (f_3) is only predictive for *hom* (AUROC 0.68). The buried amino acids showed a positive contribution to high-level production (data not shown). Results using the amino acid composition of the buried (f_4) and exposed sequence (f_5), separately, are similar to the randomized buried (f_4') and randomized exposed sequence (f_5'), indicating that neither of the two sequences is more informative than a randomly selected sequence of the same length.

Best Performance with Only Few Features

Thus far, all discussed classifiers were trained on a relatively small set of related features. Combining all features results in a large feature set which complicates both classification and interpretation. To resolve this, we used a forward feature selection protocol similar to the one used in previous work [37].

A classification performance of 0.84 AUROC was obtained for *hom* (f_{14} in Table 2), similar to the results obtained using the protein's amino acid or codon composition. Interpretation of the selected features shows a similar trend compared to the amino acid contributions observed in Figure 3A. As shown in Figure 5A, the first three selected features were almost always lysine (K), tyrosine (Y), and asparagine (N), or, as shown by a different shade of the same color, a correlated feature ($r > 0.65$).

For *het*, feature selection resulted in the best obtained prediction performance of 0.75 AUROC (f_{14}). A relatively low number of features, on average six, were selected each CV-loop, most of which were codons. Remarkably, the codon TAC (Y) was consistently selected first (Figure 5B). Methionine (M and ATG) and the codons AAC (N) and TTC (F) were most often selected second and third.

The fact that codons are selected before amino acids suggests the importance of codon usage. However, taking codon usage as features provided an AUROC of only 0.54 (f_{12}). This could be due to the heterogeneous codon usage of the different organisms in *het*. With an AUROC of 0.64, codon usage in *hom* appeared to be a little predictive.

A Role for N-glycosylation Motifs

Functional patterns, often called (short) linear motifs [43] (SLiMs), have been associated with protein targeting. The most well-known example is the C-terminal [HK]DEL motif that causes ER retention. Also a case with a secretion specific signal has been identified [44].

All proteins in our data sets contain a signal peptide, proteins with an ER-retention signal and proteins with predicted transmembrane regions are filtered out. Still, successful high-level production was observed for only half of the proteins in *hom*. Unsuccessful high-level production could for example be caused by a low production rate or a high degradation rate, resulting in a too low concentration to detect on the gel (i.e. $< 50 mg/l$). An alternative explanation could be the existence of additional retention or targeting signals. The statistical motif finding approach MEME [36] was used to search for such signals.

For *hom*, the pattern $N[GI]T$, which matches the N-glycosylation pattern $N[\hat{}P][ST]$, was found for successful high-level production. Instead of retention or targeting, this indicates importance of this post-translational modification. No other patterns related to either successful or unsuccessful high-level production were found, indicating the absence of additional generic targeting or retention signals.

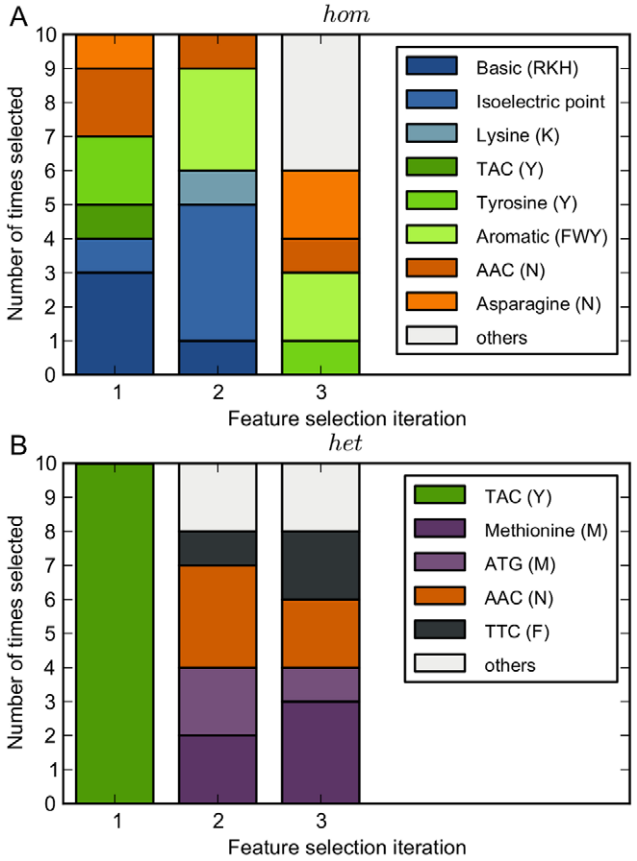

Figure 5. Feature selection. For the first three feature selection iterations (x-axis), the bar plot shows how often features were selected in the 10 CV-loops for both *hom* (**A**) and *het* (**B**). Features with a different shade of the same color are correlated ($r > 0.65$). The letters between brackets in the legend are amino acids that denote either which amino acids are in the cluster, e.g. the basic cluster contains amino acids R, K, and, H, or for which amino acid a codon encodes, e.g. codon TAC encodes for Y.
doi:10.1371/journal.pone.0045869.g005

SLiMs related to post-translational modifications can occur more than once in a sequence. Therefore we also searched for reoccurring patterns by building classifiers using fixed length pattern (k-mer) counts as features. Using the signal peptide and the protein sequence, results for $k = 2$ to $k = 5$ are shown in Table 2 (f_{15}–f_{22}). In general, classification performances rapidly drop with increasing pattern length, caused by an explosion of the number of possible k-mers that results in sparse kernels [28] which are difficult to use for classification. Again, the N-glycosylation pattern was identified. Inspection of the 3-mer classifier trained on *hom* showed that six out of the seven 3-mers with the most positive contribution match the N-glycosylation pattern.

The N-glycosylation pattern is much more abundant in *hom* than in *het*, with on average 3.37 N-glycosylation patterns per protein for *hom* compared to an average of 1.42 per protein for *het*. A clear difference is observed between the positive and negative proteins in *hom*, containing an average of 4.71 and 1.95 patterns respectively. Although much smaller, with an average of 1.72 and 1.36 patterns for the positive and negative class, respectively, *het* shows a difference as well, suggesting that the

addition of N-glycosylation sites might be useful to improve heterologous secretion [45].

Discussion

Using machine learning techniques, we explored which combinations of a large number of features best helps predicting successful high-level protein production in *A. niger*. The results show that composition-based features were most predictive, but that the exact representation – by codons, amino acids or amino acid clusters – has little influence. Taking into account predicted structural location of the amino acids did not further improve prediction results. Although all proteins have a signal peptide and the signal peptide is usually cleaved off in the ER [46], its sequence is still somewhat predictive. This suggests a role for the signal peptide in determining translocation efficiency, possibly due to a higher affinity to the SRP.

Classifiers trained on *hom* and *het* showed similar amino acid contributions, indicating that the properties found important for high-level production are generic in nature. The fact that poorer prediction performance was still obtained for *het* suggests that organism-specific properties may be important for high-level production. However, the heterogeneous nature of the *het* data and the resulting limited number of samples per donor organism hinder the identification of such properties using machine learning.

Feature selection on a larger set of features, including some derived from the sequence, confirmed that mainly composition-based features were selected in the first iterations. In fact, mainly codons and only a few amino acid features were selected for *het*. In the first three iterations, only codons were selected, implying room for production improvement by codon adaptation of heterologous proteins.

Among the composition-based features, a number of individual amino acids stood out as strongly contributing, either positively or negatively, to predicted high-level production:

- Tyrosine (Y), tryptophan (W) and phenylalanine (F) contribute positively. These aromatic amino acids are usually found in the protein core; their ability to form stacks can contribute to protein stability. A correlation between protein stability and secretion efficiency has been observed [47–49]. Moreover, improving secretion by increasing the protein stability is shown to be a successful strategy [50,51]. It is hypothesized that proteins with a high stability more frequently escape from the ER quality control system, since they will more often be in the correctly folded state, which in general is the only state to leave the ER [48,52].
- Asparagine (N) has a high positive contribution for *hom*. Since motif analysis showed the N-glycosylation pattern to be both predictive and abundant in *hom* the contribution of asparagine could be related to this post-translational process in which a specific set of enzymes catalyzes the formation of N-linked glycans. Details are still unknown, but N-linked glycans are known to play an import role in protein folding and quality control [53]. Although N-linked glycosylation is not a prerequisite for secretion [54], there is ample evidence that introduction or modification of glycosylation sites can lead to improved secretion [45,55,56].
- Methionine (M) shows a strong negative contribution. The fact that it is a sulfur-containing amino acid, and that the other sulfur-containing amino acid, cysteine (C) also has a negative contribution, suggests a negative influence of sulfur-containing amino acids. Another explanation could be that methionine is encoded by the start codon ATG, which could slow down translation due to ribosome reinitiation on alternative start sites [57].
- Lysine (K) also has a strongly negative contribution, as do the other basic amino acids arginine (R) and histidine (H) for *hom*. The positive charge, usually exposed at the protein surface, could facilitate binding to the negatively charged cell membrane, thereby preventing the protein to be filtered out, or could be related to protein thermostability due to charge-charge interactions on the protein surface [58].

In conclusion, we have exploited a large experimental dataset on production of proteins in *A. niger*, using both homologous and heterologous gene expression and employed machine learning algorithms to find combinations of features optimally predictive of presence or absence of high-level production. These features were all derived directly or indirectly from the protein sequences, and could be useful to improve industrial production rates of existing targets and to explore possibilities for new products. In future work, we intend to verify a number of the hypotheses provided here by engineering proteins to better reflect the features found to be related to high production rates.

Supporting Information

Figure S1 Shows the *hom* protein composition feature matrix (f_2). The heat map visualizes the feature matrix with the features on the *x*-axis and the proteins on the *y*-axis, the colors denote the feature value. Both the features (columns) and the proteins (rows) are clustered using complete linkage hierarchical clustering. The first bar to the right of heat map shows the protein labels, white for successful high-level production and gray for unsuccessful high-level production. The second bar to the right of the heat map shows from which donor organism the protein originates. In this case all proteins originate from *A.* niger, which is also the host organism.
(PDF)

Figure S2 Shows a heat map of the *het* protein sequence composition feature matrix (f_2), similar to Figure S1.
(PDF)

Figure S3 Shows a heat map of the *hom* codon sequence composition feature matrix (f_0), similar to Figure S1.
(PDF)

Figure S4 Shows a heat map of the *het* codon sequence composition feature matrix (f_0), similar to Figure S1.
(PDF)

Figure S5 Shows a heat map of the *hom* signal peptide composition feature matrix (f_1), similar to Figure S1.
(PDF)

Figure S6 Shows a heat map of the *het* signal peptide composition feature matrix (f_1), similar to Figure S1.
(PDF)

Figure S7 Shows a heat map of the *hom* codon usage feature matrix (f_{12}), similar to Figure S1.
(PDF)

Figure S8 Shows a heat map of the *het* codon usage feature matrix (f_{12}), similar to Figure S1.
(PDF)

Figure S9 Shows the amino acid contributions of the *hom* classifier (*x*-axis) versus amino acid costs (*y*-axis). A correlation is observed for the non-aromatic amino acids,

suggesting a preference for "cheap amino acids for high-level secretion.
(PDF)

Figure S10 Shows the amino acid contributions of the *het* classifier (*x*-axis) versus amino acid costs (*y*-axis).
(PDF)

Figure S11 Shows the classifier outcomes of the *hom* protein composition classifier. A) The histogram shows the classifier outcomes for the *hom* data set with the negatively labeled proteins in red and the positively labeled proteins in green. Note that the classifier is trained using the same data set. **B)** The histogram shows the classifier outcomes for the *A. niger* proteome in light grey. The subset of the proteome that contains a predicted signal peptide (SignalP 3.0) is shown in dark grey.
(PDF)

Table S1 Contains the labeled data set with *Aspergillus niger* proteins (*hom*) that were tested for successful high-level production and secretion. Labels pos and neg indicate successful and unsuccessful high-rate production respectively.
(XLS)

Table S2 Contains the names and abbreviations of the 14 fungal donor organisms for which there are proteins in the heterologous data set (*het*).
(TXT)

Table S3 Contains the total number of proteins and the number of successfull and unsuccessfull proteins in both *hom* and *het*, and per organism in *het*.
(XLS)

Table S4 Contains the *t*-values and corresponding *p*-values as obtained from a two-sample *t*-test on the 124 features that were used for feature selection. For both *hom* and *het*, the features are ordered by the absolute *t*-value.
(XLS)

Table S5 Contains the selected features in each CV-loop with forward feature selection on both *hom* and *het*.
(XLS)

Table S6 Contains the best performing amino acid clusters in each CV-loop obtained with the amino acid clustering procedure on both *hom* and *het*.
(TXT)

Author Contributions

Conceived and designed the experiments: BAB DR MJTR JAR MH HJP LW. Performed the experiments: BAB. Analyzed the data: BAB DR MJTR JAR. Contributed reagents/materials/analysis tools: BAB MH. Wrote the paper: BAB DR JAR MJTR.

References

1. Lubertozzi D, Keasling JD (2009) Developing *Aspergillus* as a host for heterologous expression. Biotechnology Advances 27: 53–75.
2. Pel HJ, de Winde JH, Archer DB, Dyer PS, Hofmann G, et al. (2007) Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88. Nature Biotechnology 25: 221–231.
3. Andersen MR, Salazar MP, Schaap PJ, van de Vondervoort PJI, Culley D, et al. (2011) Comparative genomics of citric-acid-producing *Aspergillus niger* ATCC 1015 versus enzyme-producing CBS 513.88. Genome Research 21: 885–897.
4. Punt PJ, van Biezen N, Conesa A, Albers A, Mangnus J, et al. (2002) Filamentous fungi as cell factories for heterologous protein production. Trends in Biotechnology 20: 200–206.
5. Nevalainen KM, Te'o VS, Bergquist PL (2005) Heterologous protein expression in filamentous fungi. Trends in Biotechnology 23: 468–474.
6. Fleißner A, Dersch P (2010) Expression and export: recombinant protein production systems for *Aspergillus*. Applied Microbiology and Biotechnology 87: 1255–1270.
7. Gouka RJ, Punt PJ, van den Hondel CAMMJ (1997) Efficient production of secreted proteins by *aspergillus*: progress, limitations and prospects. Applied Microbiology and Biotechnology 47: 1–11.
8. Conesa A, Punt PJ, van Luijk N, van den Hondel CAMJJ (2001) The secretion pathway in filamentous fungi: a biotechnological view. Fungal Genetics and Biology 33: 155–171.
9. Archer DB, Peberdy JF (1997) The molecular biology of secreted enzyme production by fungi. Critical Reviews in Biotechnology 17: 273–306.
10. Carvalho NDSP, Arentshorst M, Kooistra R, Stam H, Sagt CM, et al. (2011) Effects of a defective ERAD pathway on growth and heterologous protein production in *Aspergillus niger*. Applied Microbiology and Biotechnology 89: 357–373.
11. Jacobs DI, Olsthoorn MA, Maillet I, Akeroyd M, Breestraat S, et al. (2009) Effective lead selection for improved protein production in *Aspergillus niger* based on integrated genomics. Fungal Genetics and Biology 46: S141– S152.
12. Guillemette T, van Peij NNME, Goosen T, Lanthaler K, Robson GD, et al. (2007) Genomic analysis of the secretion stress response in the enzyme-producing cell factory *Aspergillus niger*. BMC Genomics 8: 158.
13. Plotkin JB, Kudla G (2010) Synonymous but not the same: the causes and consequences of codon bias. Nature Reviews Genetics 12: 32–42.
14. Horton P, Park KJ, Obayashi T, Nakai K (2006) Protein subcellular localization prediction with WoLF PSORT. In: Proceedings of the 4th annual Asia Pacific bioinformatics conference APBC06, Taipei, Taiwan. Citeseer, volume 39, p.48.
15. Pierleoni A, Martelli PL, Fariselli P, Casadio R (2006) BaCelLo: a balanced subcellular localization predictor. Bioinformatics 22: e408–e416.
16. Blum T, Briesemeister S, Kohlbacher O (2009) MultiLoc2: integrating phylogeny and gene ontology terms improves subcellular protein localization prediction. BMC Bioinformatics 10: 274.
17. Magnan CN, Randall A, Baldi P (2009) SOLpro: accurate sequence-based prediction of protein solubility. Bioinformatics 25: 2200.
18. Sonnenburg S, Zien A, Philips P, Rätsch G (2008) POIMs: positional oligomer importance matricesunderstanding support vector machine-based signal detectors. Bioinformatics 24: i6–i14.
19. Briesemeister S, Rahnenführer J, Kohlbacher O (2010) Going from where to why – interpretable prediction of protein subcellular localization. Bioinformatics 26: 1232–1238.
20. van Dijck PWM, Selten G, Hempenius RA (2003) On the safety of a new generation of dsm aspergillus niger enzyme production strains. Regulatory Toxicology and Pharmacology 38: 27–35.
21. Dyrløv Bendtsen J, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. Journal of Molecular Biology 340: 783–795.
22. Krogh A, Larsson BÈ, Von Heijne G, Sonnhammer ELL (2001) Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. Journal of Molecular Biology 305: 567–580.
23. Käll L, Krogh A, Sonnhammer ELL (2004) A combined transmembrane topology and signal peptide prediction method. Journal of Molecular Biology 338: 1027–1036.
24. Dondoshansky I (2002) Blastclust (NCBI Software Development Toolkit). NCBI, Bethesda, Md.
25. Petersen B, Petersen TN, Andersen P, Nielsen M, Lundegaard C (2009) A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Structural Biology 9: 51.
26. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology 292: 195–202.
27. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2: 27.
28. Ben-Hur A, Ong CS, Sonnenburg S, Scholkopf B, Rätsch G (2008) Support vector machines and kernels for computational biology. PLoS Computational Biology 4: e1000173.
29. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognition Letters 27: 861–874.
30. Taylor WR (1986) The classification of amino acid conservation. Journal of Theoretical Biology 119: 205–218.
31. Sharp PM, Li WH (1987) The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research 15: 1281.
32. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, et al. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25: 1422.
33. Jones E, Oliphant T, Peterson P (2001). SciPy: Open source scientific tools for Python. Available: http://www.scipy.org/.
34. Leslie C, Eskin E, Noble WS (2002) The spectrum kernel: a string kernel for SVM protein classification. In: Pacific Symposium on Biocomputing. Hawaii, USA., volume 575, 564–575.

35. Sonnenburg S, Rätsch G, Henschel S, Widmer C, Behr J, et al. (2010) The SHOGUN machine learning toolbox. The Journal of Machine Learning Research 99: 1799–1802.
36. Bailey TL, Williams N, Misleh C, Li WW (2006) Meme: discovering and analyzing dna and protein sequence motifs. Nucleic Acids Research 34: W369–W373.
37. van den Berg BA, Nijkamp JF, Reinders MJT, Wu L, Pel HJ, et al. (2010) Sequence-based prediction of protein secretion success in Aspergillus niger. In: Proceedings of Pattern Recegnition in Bioinformatics 2010. Springer, 3–14.
38. Nakashima H, Nishikawa K (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. Journal of Molecular Biology 238: 54–61.
39. Cedano J, Aloy P, Perez-Pons JA, Querol E (1997) Relation between amino acid composition and cellular location of proteins. Journal of Molecular Biology 266: 594–600.
40. Rätsch G, Sonnenburg S (2004) Kernel Methods in Computational Biology. The MIT Press, 277 p.
41. Toussaint N, Widmer C, Kohlbacher O, Rätsch G (2010) Exploiting physicochemical properties in string kernels. BMC Bioinformatics 11: S7.
42. Akashi H, Gojobori T (2002) Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. Proceedings of the National Academy of Sciences 99: 3695.
43. Neduva V, Russell RB (2005) Linear motifs: evolutionary interaction switches. FEBS letters 579: 3342–3345.
44. Hiller NL, Bhattacharjee S, van Ooij C, Liolios K, Harrison T, et al. (2004) A host-targeting signal in virulence proteins reveals a secretome in malarial infection. Science 306: 1934.
45. Sagt C, Kleizen B, Verwaal R, De Jong M, Muller W, et al. (2000) Introduction of an Nglycosylation site increases secretion of heterologous proteins in yeasts. Applied and Environmental Microbiology 66: 4940.
46. von Heijne G (1990) The signal peptide. Journal of Membrane Biology 115: 195–201.
47. Kowalski JM, Parekh RN, Wittrup KD (1998) Secretion efficiency in saccharomyces cerevisiae of bovine pancreatic trypsin inhibitor mutants lacking disulfide bonds is correlated with thermodynamic stability. Biochemistry 37: 1264–1273.
48. Kowalski JM, Parekh RN, Mao J, Wittrup KD (1998) Protein folding stability can determine the efficiency of escape from endoplasmic reticulum quality control. Journal of Biological Chemistry 273: 19453.
49. Whyteside G, Alcocer MJC, Kumita JR, Dobson CM, Lazarou M, et al. (2011) Native-state stability determines the extent of degradation relative to secretion of protein variants from Pichia pastoris. PLoS ONE 6: e22692.
50. Shusta EV, Kieke MC, Parke E, Kranz DM, Wittrup KD (1999) Yeast polypeptide fusion surface display levels predict thermal stability and soluble secretion efficiency1. Journal of Molecular Biology 292: 949–956.
51. Kjeldsen T, Ludvigsen S, Diers I, Balschmidt P, Sørensen AR, et al. (2002) Engineering-enhanced protein secretory expression in yeast with application to insulin. Journal of Biological Chemistry 277: 18245–18248.
52. Ellgaard L, Helenius A (2003) Quality control in the endoplasmic reticulum. Nature Reviews Molecular Cell Biology 4: 181–191.
53. Helenius A, M A (2001) Intracellular functions of N-linked glycans. Science 291: 2364.
54. Eriksen SH, Jensen B, Olsen J (1998) Effect of N-linked glycosylation on secretion, activity, and stability of α-amylase from *Aspergillus oryzae*. Current Microbiology 37: 117–122.
55. van den Brink HJM, Petersen SG, Rahbek-Nielsen H, Hellmuth K, Harboe M (2006) Increased production of chymosin by glycosylation. Journal of Biotechnology 125: 304–310.
56. Liu Y, Nguyen A, Wolfert RL, Zhuo S (2009) Enhancing the secretion of recombinant proteins by engineering N-glycosylation sites. Biotechnology Progress 25: 1468–1475.
57. Kochetov AV (2008) Alternative translation start sites and hidden coding potential of eukaryotic mRNAs. Bioessays 30: 683–691.
58. Strickler SS, Gribenko AV, Gribenko AV, Keiffer TR, Tomlinson J, et al. (2006) Protein stability and surface electrostatics: a charged relationship. Biochemistry 45: 2761–2766.