1 Ionization Efficiency (IE) prediction.

1.1 Notations

Let $peptide_i$ be a product of digestion of $protein_j$.

I_i : Intensity of $peptide_i$

IE_i : Ionization Efficiency of $peptide_i$

c_i : concentration of $peptide_i$ after trypsin digestion

c_j : concentration of $protein_j$ before trypsin digestion

n : total number of detected peptides

1.2 Model

By definition, $I_i = IE_i c_i$. Assuming a perfectly efficient trypsin digestion, if $peptide_i$ is tryptic, then $c_i = c_j$. Thus, for any peptide i belonging to protein j :

$$\log_{10}(I_i) = \log_{10}(IE_i) + \log_{10}(c_j)$$

Let $y_i = \log_{10}(I_i)$, and $b_j = \log_{10}(c_j)$.

$$y_i = \log_{10}(IE_i) + b_j$$

We model $\log_{10}(IE_i)$ as a linear combination of a K-long feature vector $X_{i,k}$ derived from $peptide_i$'s amino acid sequence, weighted with weights $W = [w_1, ..., w_K]$

$$\log_{10}(IE_i) = \sum_k (w_k X_{i,k}) + \epsilon_i$$
$$y_i = \sum_k (w_k X_{i,k}) + b_j + \epsilon_i$$

where $\epsilon_i \sim \mathcal{N}(\mu = 0, \sigma^2)$, and ϵ_i is independent of X. The equation resembles that of a Gaussian-noise simple linear regression model, except for the fact that there is a separate intercept b_j per protein. The likelihood of the system is given by:

$$\begin{aligned}
\mathcal{L} &= \prod_j \prod_{i \in protein_j} p(y_i | x_i; W, B, \sigma^2) \\
&= \prod_j \prod_{i \in protein_j} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - (b_j + \sum_k w_k X_{i,k}))^2}{2\sigma^2}} \\
\log(\mathcal{L}) &= \frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2} \sum_j \sum_{i \in protein_j} (y_i - b_j - \sum_k w_k X_{i,k})^2
\end{aligned}$$

1.3 Data preparation

We selected unmodified peptides of charge 2 to train our model. Using the MQ output file 'evidence.txt' as our input. Peptides belonging to proteins detected by only one peptide were excluded from the analysis. The remaining peptides were grouped by sequence, and their intensity was defined as the sum of their intensities across all fractions.
For each peptide, we then computed features capturing their amino acid composition, their length, and potential mis-cleavage issues (see Table 1)

Features were then centered and reduced to unit variance to create the feature matrix X

1.4 Parameter fitting and results

We obtained maximum likelihood estimates for W, B and σ by maximizing the log likelihood described in section 1.2., using the L-BFGS-B implementation of Python's sklearn module. Figure 1 shows a good agreement between the predicted $\log_{10}(IE)$ (defined as $\sum_k(w_k X_{i,k})$) and the observed $\log_{10}(IE)$ (defined as $y_i - b_j$), with a Pearson correlation coefficient R = 0.67. In order to control for over-fitting, we further divided the set of proteins in two equally sized groups, and fitted the weights W and $\log_{10}(protein\,levels)$ B separately for both groups. We found a very good agreement between the W vectors in both groups, confirming that we were not over-fitting the data (Figure 2). We report the value of the fitted W coefficients in Figure 3

Table 1: Features computed for this analysis

Name	Description	Length
$count_{AA}$	# of occurences of AA in peptide	20
$count_{RP}$	# of occurences of the subsequence 'RP' in peptide	1
$count_{KP}$	# of occurences of the subsequence 'KP' in peptide	1
$N_{term}\,Pro$	1 if peptide starts with Pro, 0 otherwise	1
-2 is R	1 if the aa in position -2 relative to the N_{term} cleavage site is 'R', 0 otherwise	1
-2 is K	1 if the aa in position -2 relative to the N_{term} cleavage site is 'K', 0 otherwise	1
-1 is R	1 if the aa in position -1 relative to the N_{term} cleavage site is 'R', 0 otherwise	1
-1 is K	1 if the aa in position -1 relative to the N_{term} cleavage site is 'K', 0 otherwise	1
+1 is R	1 if the aa in position +1 relative to the C_{term} cleavage site is 'R', 0 otherwise	1
+1 is K	1 if the aa in position +1 relative to the C_{term} cleavage site is 'K', 0 otherwise	1
+1 is P	1 if the aa in position +1 relative to the C_{term} cleavage site is 'P', 0 otherwise	1
inverse length	inverse of the peptide's length	1
length	length of the peptide	1
	Total	32

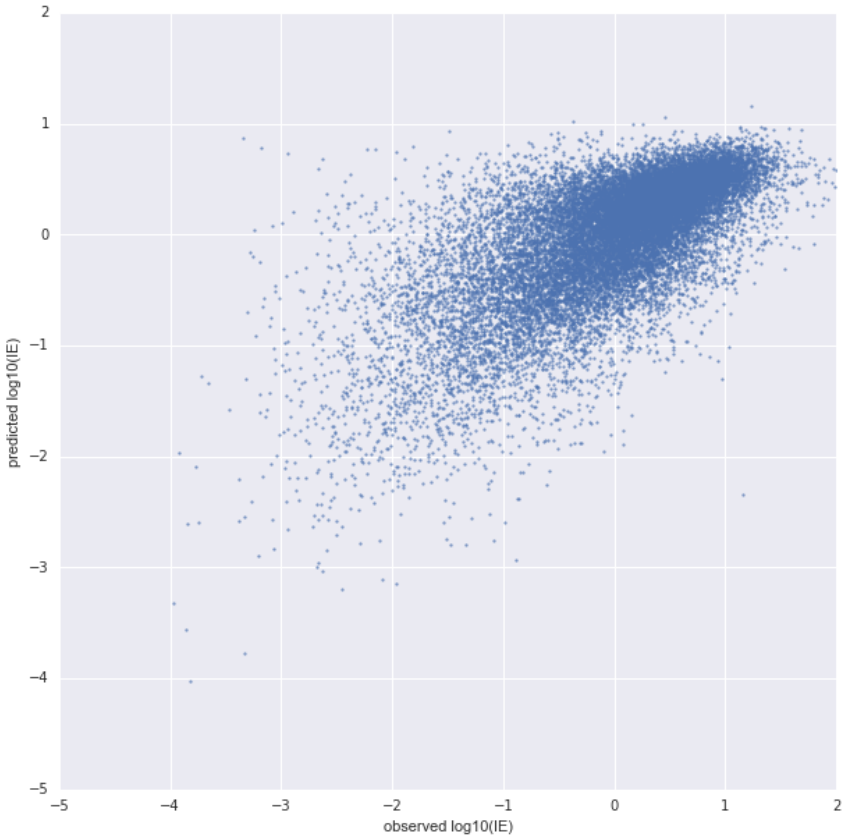

Figure 1: The predicted $\log_{10}(IE)$ was computed as $\sum_k(w_k X_{i,k})$, and the observed $\log_{10}(IE)$ was defined as $y_i - b_j$. Pearson correlation coefficient = 0.69, $\sigma = 0.58$

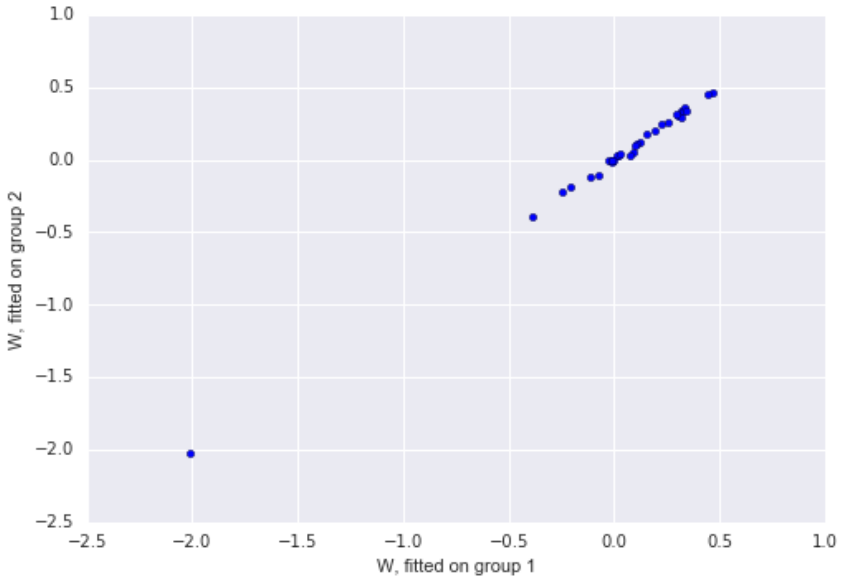

Figure 2: Weights of the regression coefficients W, fitted separately on each half of the dataset, are plotted against one another. Pearson correlation coefficient > 0.999

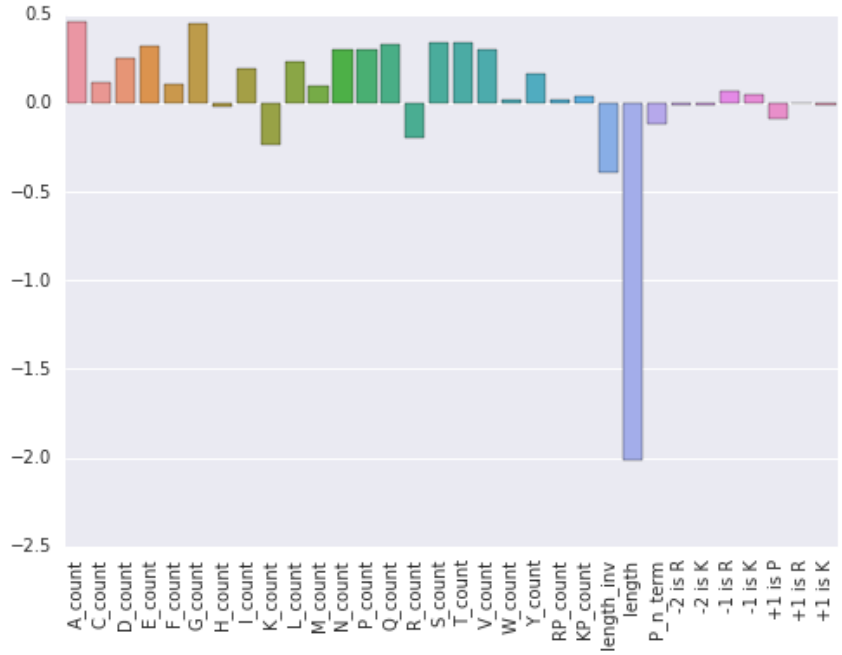

Figure 3: Weights of the regression coefficients W, fitted on the entire dataset. The weigths are given in unit of $\log_{10}(IE)$