1 Definitions

1. peptide - Amino acid sequence and modification masses. For a given peptide p let $p = ((a_1, z_1), \cdots (a_n, z_n))$ where a represents an amino acid and z represents a modification.

2. variant - combination of modification mass an location for a particular peptide. For a variant v of peptide p let $v = ((m_1, x_1), \cdots, (m_n, x_n))$ where m represents the modification mass and x represents the position within peptide p.

3. a variant of a peptide is an assignment of modification masses to specific amino acid positions (i.e., *sites*) on a peptide sequence. At most one modification mass is allowed on each amino acid; multiply-modified amino acids are modeled using modifications of the resulting aggregate mass. For example, di-methylation is represented using nominal mass 28 Da instead of modeled as two methylations, each of mass 14 Da.

4. Spectrum - vector of mass intensity pairs. For a given spectrum $\vec{S}$ let $\vec{S} = ((m_1, y_1), \ldots, (m_n, y_n))$ where m_i and y_i represent the mass and intensity of *peak* i.

5. Peptide masses - for simplicity, we define theoretical peptide masses $Masses(P)$ as a set of prefix (N-terminal) and suffix (C-terminal) fragment masses, where each fragment mass is determined by the sum of its constituent amino acid masses. In reality, our method is implemented using the common types of fragments resulting from Collision Induced Dissociation (for our purposes, b and y single and doubly charged and b and y isomers with an extra dalton) and can be easily configured to support other types of mass spectrometry peptide dissociation strategies.

2 Enumeration of Modification Variants

1. Given a peptide identification of length n and consisting of the multiset of modification masses $\mathcal{M} = m_1 \cdots m_k$ with the number of duplicate masses equalling d.

2. Generate all distinct permutations of length n consisting of $k + 1$ elements (k modifications with a placeholder element for unmodified positions) to generate $\binom{n}{k} \cdot \frac{k!}{d!}$ variants.

3 Peptide variant spectra

3.1 Similarity between spectra of modified and unmodified peptide variants

1. Given a spectrum S from a peptide P, we define $Intensities(S, P)$ as the intensities of the spectrum peaks in S at $Masses(P)$. Without loss of generality, the vector $Intensities(S, P)$ is always normalized to Euclidian norm 1.

2. Given a spectrum S from a peptide P and a spectrum S_v from a modified peptide variant P_v, we define the similarity between the extracted peak intensities as

$$\begin{aligned} Similarity(S, S_v) &= \cos(Intensities(S, P), Intensities(S_v, P_v)) \\ &= Intensities(S, P) \cdot Intensities(S_v, P_v) \end{aligned}$$

Supplementary Materials Section **??** offers an in detail look at how the intensities of modified spectra compare to unmodified spectra from the same peptide sequence.

3.2 Prediction from unmodified peptide spectra

First, we choose the best candidate unmodified spectrum:

1. Given a set of spectra $S_1 \cdots S_n$ all from peptide P

2. Given a modified spectrum S' from modified peptide P' whose unmodified version is P

3. Choose the spectrum with the highest $Similarity(S_i, S')$

Note that there are other methods for choosing can best candidate, for example, giving preference to unmodified spectra from the same dataset as the modified spectrum.

1. Given a spectrum S from an unmodified peptide P, we want to predict a spectrum for P_v, a modified variant of P.

2. Extract peak intensities from S at $Masses(P)$ and use these to set the corresponding peak intensities in S_v at $Masses(P_v)$.

4 Linear programming model

4.1 what is being solved

An experimental spectrum is modeled as a linear combination of the intensity vectors for all possible variants of the same modified peptide sequence.

Inputs:

1. A spectrum S from peptide P.
2. A spectrum S' from modified peptide P' whose unmodified sequence matches P
3. A set of variants of P', V
4. Peak tolerance δ

Generate a vector T of all possible ion masses in variants.

1. Take $Masses(P)$ and add them to T.
2. Take multiset of modification masses $\mathcal{M} = \{m_1, \ldots, m_k\}$ from P' and generate all combinations of mods $\binom{M}{1} \cdots \binom{M}{k}$. Sum the masses contained in each combination to get the set of modification mass shifts C.
3. For each mass t_i in T, add modification shifts $C = c_1 \cdots c_j$, $t_i + c_j = t_i j$. Add $t_{i1} \cdots t_{ij}$ to set T.

Generate an LP where the observed intensity in the modified spectrum is expected to be a summation of the expected intensities of each ion scaled by the abundance of each variant.

1. For each theoretical ion t_j in T, if a peak from S' is within tolerance δ, then observed peak intensity $O_j = y$ otherwise observed peak intensity $O_j = 0$
2. For each theoretical ion t_j if there is a variant v_i where $t_j \in Masses(v_i)$ in the expected peak tolerance, assume that Q_i is contributing to overall intensity O_j. We add all such variants to $\mathcal{V}_j \subset V$ To generate expected intensity of the peak, we take the mass unmodified version of the ion of t_j from S to get d_j.

3. For each theoretical ion, we assume that the observed peak is the sum of the contribution of all variants scaled by their quantity and expected intensity. :

$$O_j \approx \sum_{i=1}^{|\mathcal{V}_j|} d_j \times Q_i$$

From this, we are able to approximate the error for each peak as follows

$$\epsilon_j = O_j - \sum_{i=1}^{|\mathcal{V}_j|} d_j \times Q_i$$

We then generate an LP which minimizes the error of each peak:

Input	Output	Formulation
d_j for every ion j O_j for every ion j	Q_i for every variant v_i	$\min \quad \sum_{j=1}^{r} \lvert\varepsilon_j\rvert$ s.t. $\quad \varepsilon_j = O_j - \sum_{i=1}^{\lvert\mathcal{V}_j\rvert} d_j \times Q_i$ $\quad Q_i \geq 0$ $Q_i s$ are normalized prior to output so that $\sum_i (Q_i) = 1$

5 Grouping procedure

Inputs

1. Variants $v_1 \cdots v_n$ and quantities $Q_1 \cdots Q_n$ from a single peptide where the quantity of v_i is represented by Q_i.
2. Modified spectrum S'
3. Grouping threshold λ

For a variant group g_i consisting of $v_1 \cdots v_n$, $Masses(g_i) = Masses(v_1) \cup Masses(v_2) \cdots \cup Masses(v_n)$.

1. Form n variant groups containing a single variant, $g_1 \cdots g_n$.

2. Find the distance between pairs of groups. To calculate the distinguishing intensity between g_i and g_j, take $Masses(Difference) = Masses(g_i) \triangle Masses(g_j)$ and sum the peak intensities from S' which match $Masses(Difference)$ within our peak tolerance to get the distinguishing intensity. We then calculate the distance by dividing the sum of distinguishing intensity by the total identified intensity.

3. Find the two variant groups g_i and g_j with the lowest distance. If these groups have a distance above λ, stop.

4. Merge the two closest groups and create new group $\vec{g_k}$. Q_k is defined as the sum of Q_i and Q_j

5. Recompute distances between g_k all other groups.

6 Calculating cosine for modified vs. theoretical spectra

Inputs

1. Variant groups $g_1 \cdots g_n$ with quantities $Q_1 \cdots Q_n$ from a single peptide where the quantity of g_i is represented by Q_i

2. A spectrum S from peptide P.

3. A spectrum S' from modified peptide P' whose unmodified sequence matches P

Generate theoretical spectra for all variant groups.

1. For each cluster g_i, extract peak intensities from S at $Masses(P)$ and use those to set the corresponding expected intensities S_i at $Masses(g_i)$.

2. Scale each peak in S_i by quantity indicated by Q_i. Add all peaks to theoretical spectrum T. Sum intensities of any matching peaks already in T.

Calculate cosine between theoretical and modified spectrum

1. Calculate $Similarity(S', T)$ to generate the theoretical cosine.

7 FLR

7.1 Determining decoys

Inputs

1. Variant group c consisting of peptide variants.
2. Vector of modification masses and their expected amino acid residues $E = \{(m_1, a_1) \cdots (m_n, a_n)\}$

Determine whether peptide variant is a decoy

$$isDecoy(v) = \begin{cases} 0 & \text{if for each amino acid residue non-zero modification pair, } a_i, z_i \in E \\ 1 & \text{otherwise} \end{cases}$$

Determine whether a variant group is a decoy

$$isDecoy(g) = \begin{cases} 0 & \text{if there exists a peptide variant } v_i \in c \text{ where } isDecoy(v_i) = 0 \\ 1 & \text{otherwise} \end{cases}$$

7.2 Calculate decoy scaling factor

1. Take all variant groups G for all peptides
2. $TP = n - \sum_{i=1}^{|G|} isDecoy(g_i)$, $FP = \sum_{i=1}^{|G|} isDecoy(g_i)$.
3. Scaling factor $\rho = \frac{TP}{FP}$

7.3 Calculate FLR non-successive thresholds

Inputs

1. Variant groups $g_1 \cdots g_n$ with associated quantities $Q_1 \cdots Q_n$ and associated theoretical cosine $t_1 \cdots t_n$.
2. FLR cutoff γ

Calculate FLR

1. Sort variant groups by quantity and theoretical cosine.
2. Count number of decoy hits I and number of target hits T at each index
3. Find maximum index m for $\frac{I*\rho}{T} \leq \gamma$
4. Return variant groups with indices lower than m.

7.4 Calculate FLR successive thresholds

Inputs

1. Variant group sets $G_1 \cdots G_n$ where the grouping threshold of G_i is equivalent to grouping threshold i.
2. FLR cutoff γ

Calculate FLR

1. Starting at the lowest threshold 1, sort G_1 according to its associated quantities and theoretical cosines.
2. Filter G_1 by FLR as described in Section7.3 to get G'_1
3. For $G_2 \cdots G_n$ if there is a peptide spectrum match $p_i \in G'_1$ filter all associated variant groups for that peptide spectrum match for that group.
4. Repeat with the next lowest threshold
5. Once all thresholds have been run, return $G'_1 \cdots G'_n$