## Introduction

The number of peptides which could potentially be available for presentation in the immune response greatly increases when peptide splicing and post-translational modification is considered. This poses a challenge for researchers; if PTMs and spliced peptides are considered, peptidomic databases become too large to for computational software to effectively process. Alternatively, if these additional peptides are ignored salient peptides may be missed. In response to this problem, we have created a software package MersSplicer which generates spliced and modified peptide databases that have been filtered via comparison with tandem mass spectrometry (MS/MS) data. We have also designed a workflow utilising this software which improves peptide recognition while ensuring spliced and modified peptides are not ignored.

## Workflow

Workflow summary: need flow diagram for this also.

Following of origin data.

Tandem mass spectrometry (MS/MS) is used to generate an MGF file containing the complete set of spectra in a peptidome of interest.

THE MERSSPLICER PROGRAM

The generated MGF file and a FASTA file containing a series of proteins are then input into the MersSplicer software. The software firstly generates all the peptides which can be formed from the proteins in the input FASTA file, subject to the splice type and criteria input by the user. Other than selecting to perform linear, cis or trans splicing, the user can select the minimum and maximum length of peptides, up to six PTMs which are to be applied to the generated peptides and the maximum number of modifications which can be made to a single peptide.

After the peptides are generated the software begins compiling the data necessary to compare potential peptides to the spectra within the MGF file. The monoisotopic mass of each peptide is calculated and converted to mass-to-charge ratios for each of the charge states selected by the user for consideration. The masses of the b and y ions of each peptide are similarly calculated and stored for comparison against the MS/MS spectra.

The next stage of the program compares the data formulated for each peptide to the spectra in the MGF file. Peptides which are considered a match as per the user input criteria are written to an output FASTA file, while those which do not match are ignored. The filtering criteria which the user can manipulate includes the required precursor accuracy in PPM, the minimum b and y ion accuracy in Dalton and the minimum percentage of b and y ions which must be matched.

In the filtering stage, the mass to charge ratio of each peptide is compared to the precursor masses for spectra of the same charge state. If no precursor mass exists which is within the required input precursor accuracy, the peptide is ignored. For every spectrum which meets the precursor accuracy criteria, the program proceeds to compare the b and y ion masses of the peptide to other peaks in the spectra. A b or y ion is considered to have matched if its mass and the charge to mass ratio of one of the fragment ions are within the b and y ion accuracy. If the percentage of b and y ions in a peptide which matched exceeds the input b and y ion percentage requirement, the peptide is considered to have matched and is written to an output FASTA file. If not enough b and y ions match, the peptide is again ignored.

PEPTIDE CONCATENATION

The next stage of the workflow is to use the formulated database as an input to the PEAKS software. However, the PEAKS software has a limited input size of about 40,000 amino-acid sequences. The output databases being generated by the MersSplicer software are often in excess of one million peptides and thus cannot be used directly as an input into PEAKS. A simple solution is to concatenate the peptides in the database head to tail to create fewer but longer amino acid sequences. The problem with this approach is that it creates unwanted inputs at the border between each peptide. PEAKS essentially splits input sequences into linear spliced peptides before conducting its analysis, meaning it will analyse a huge number of peptides which were not in the original database if end to tail concatenation is used.

The extent of this issue is reduced via an algorithmic approach to concatenation which is utilised in the MersSplicer software. The algorithm firstly deletes all peptides from the database which exist as sub-sequences of longer peptides. That is, the peptide MVDREQLV would be deleted if the peptide MVDREQLVRTS also exists in the database. A version of the binary search algorithm is then used to combine peptide sequences together which have and identical prefix and suffix. That is, MVDREQLV could be combined with DREQLVRTS to create a new peptide MVDREQLVRTS. This form of concatenation is repeated until there are no matching prefixes and suffixes at all, at which point the sequences are simply concatenated head to tail to achieve the desired number of sequences of approximately four thousand. The algorithm combines the longest matching prefixes and suffixes first to minimise the additional peptides created.

This “smart concatenation” is run after the initial filtered database has been compiled and output to a FASTA file. There is a checkbox on the software interface enabling the user to automatically produce a concatenated file. Alternatively, the concatenation script can be run separately.