**Introduction**

- Current metaproteomics algorithms limited by speed and sensitivity.
- Loss of sensitivity can reduce the number of identified peptides by as much as 50% (Fig 1).

**Figure 1.** A GeoBacter MS/MS dataset is searched against multiple protein sequence databases. The loss of identified peptides is solely attributable to the increased database size. For a metagenomics library of 1-10Gb, a 50% loss in peptides is predicted.

**Spectral Networks**

- A spectral network is a cluster of related pairwise spectrum alignments. (Fig 2)
- Identifies similar spectra without requiring: identifications a priori
- Creates dense networks of related spectra.
- Peptides from clustered spectra share sequence similarity.

**Figure 2 – Spectral Network.** Three spectra aligned to each other in a network are shown. The peptide sequences of the spectra are polymorphisms of each other.

**AlignGF**

- Current algorithms for spectrum alignment calculate p-values based on bulk statistics, not individual spectra.
- AlignGF calculates rigorous p-values (Fig 3), using a probability density function of all possible scores, where matching a peak is considered as an independent Bernoulli event with probability θ.
- AlignGF validates pairwise Spectrum Alignments by calculating the probability of the peak matches for each spectrum (Fig 4).

**Figure 3 – AlignGF.** All possible subsets of a peak are enumerated and their score tabulated (left side). The frequency of achieving a given score is transformed into the probability density function.

**Figure 4 – Validating Alignments.** After aligning spectra, the score of picked peaks is compared against the PDF (Figure 3) to access statistical significance.

**Tag Filtering**

- Time spent in alignment grows with the number of spectra, O(n^2).
- Tagging identifies spectra likely to share short amino acid substrings (Fig 5, 6).
- Filtering alignment pairs via tagging reduces the number of comparisons 200 fold.
- Tag filtering improves the overall run time >200x.

**Figure 5 – Sequence tag.** A tag is a short partial interpretation of a spectrum

**Figure 6 – Tag-based filtering.** Most ‘alignable’ spectra share short sequence tags. 50 tags of length 3 are generated per spectrum. Spectral pairing requires at least one matching tag between two spectra.

**Results**

- We use the AlignGF approach to validate pairwise alignments in spectral networks.
- Datasets show the performance of Spectral Networks to correctly identify spectrum pairs whose peptides are related, e.g. sequences from different organisms.
- In addition to increased sensitivity and precision (Fig 7), AlignGF also validates alignments to unexplained spectra (novel polymorphisms).

**Figure 7 – Performance comparison.** Spectral pairing sensitivity and precision comparison between AlignGF and old process procedures. This graph plots all alignment pairs. Sensitivity is the percent of correct pairs recovered. Precision is the percent of total pairs that are correct.

**Table 1 – Performance comparison.** The sensitivity and precision of several different methods for accessing significance of spectral alignments. Distance 0 pairs are those with identical peptide sequences. Distance 1 pairs have one amino acid polymorphism. Prefix/suffix pairs are those where one peptide sequence is longer (a substring of the other).