Algorithms Enabling Metaproteomics

Seungjin Na[1], Stephen Callister[2], Nuno Bandeira[1], Samuel H Payne[2]
1 University of California, San Diego, La Jolla CA; 2 Pacific Northwest National Laboratory, Richland WA

Pacific Northwest NATIONAL LABORATORY

Abstract

- Metaproteomics is the analysis of proteins from microbial communities.
- Metaproteomics can often reveal the dynamic relationships between microbes and their environment.
- Current algorithms and analysis protocols have inherent sensitivity limitations.
- Our objective is to dramatically improve biological insights by improving the sensitivity via novel computational algorithms.

Project Goals

- Metagenomic sequencing does not reveal active metabolic or signaling processes.
- Metaproteomics experiments contain proteins from many species (10s-1000s).
- Protein sequences share noticeable similarity due to evolution.
- Spectral networks link spectra of related sequences, peptides of the same homolog from different organisms.
- Linking spectra is done through spectrum alignment.
- Aligned spectra allow for the transfer of sequence from one spectrum onto an adjoining spectrum.

Conclusions

- AlignGF dramatically improves the sensitivity of spectrum alignment
- Tag filtering provides a >200x speed up in SpectralNetwork creation
- AlignGF dramatically expands the spectral network while simultaneously improving alignment quality.

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)

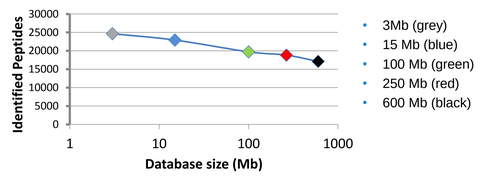

Figure 1. A *Geobacter* MS/MS dataset is searched against multiple protein sequence databases. The loss of identified peptides is solely attributable to the increased database size. For a metagenomics library of 1-10Gb, a 50% loss in peptides is predicted.

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.

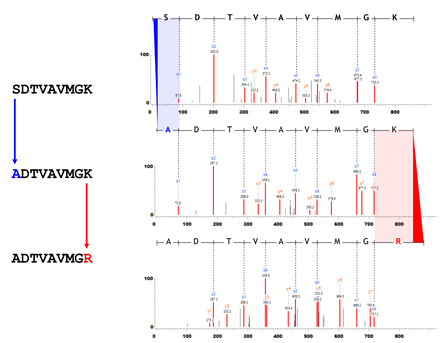

Figure 2 – Spectral Network. Three spectra aligned to each other in a network are shown. The peptide sequences of the spectra are polymorphisms of each other.

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).

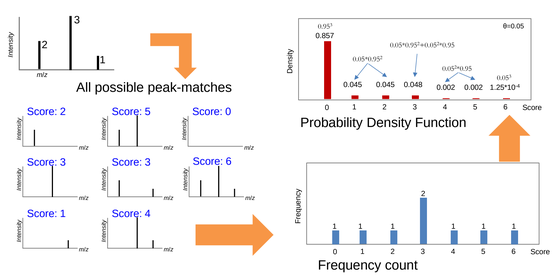

Figure 3 – AlignGF. All possible subsets of a peak are enumerated and their score tabulated (left side). The frequency of achieving a given score is transformed into the probability density function.

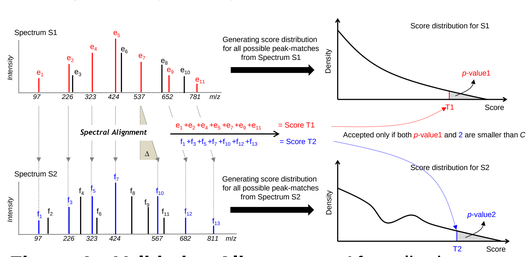

Figure 4 – Validating Alignments. After aligning spectra, the score of picked peaks is compared against the PDF (Figure 3) to access statistical significance.

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.

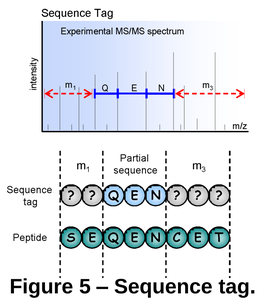

Figure 5 – Sequence tag. A tag is a short partial interpretation of a spectrum

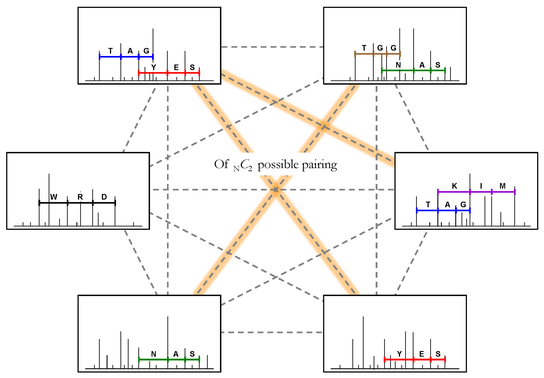

Figure 6 – Tag-based filtering. Most 'alignable' spectra share short sequence tags. 50 tags of length 3 are generated per spectrum. Spectral pairing requires at least one matching tag between two spectra.

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 with a polymorphism.
- In addition to increased sensitivity and precision (Fig 7), AlignGF also validates alignments to unexplained spectra (novel polymorphisms).

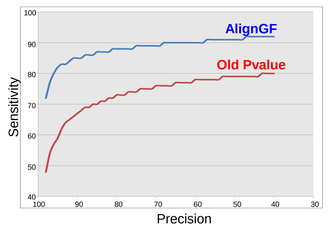

Figure 7 – Performance comparison . Spectral pairing sensitivity and precision comparison between AlignGF and old Pvalue procedures. This graph plots all alignment pairs. Sensitivity is the percent of correct pairs recovered. Precision is the percent of total pairs that are correct.

			Old Pvalue		AlignGF	
	TRUE pairs	FALSE pairs	sensitivity	precision	sensitivity	precision
distance 0	5,066	63,785	72.5%	98.8%	87.8%	98.3%
distance 1	2,585	4,528,129	64.4%	88.2%	72.3%	94.2%
prefix/suffix (0.9 overlap)	2,265	10,591,425	63.9%	80.1%	74.0%	91.2%
prefix/suffix (0.8 overlap)	1,098	9,965,794	59.7%	66.3%	77.3%	90.7%

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).

U.S. DEPARTMENT OF ENERGY

Acknowledgements
This research was funded and Early Career award to SHP from the U. S. Department of Energy (DOE) Biological and Environmental Research. Work was performed in the Environmental Molecular Sciences Laboratory (EMSL), a DOE/BER national scientific user facility at Pacific Northwest National Laboratory.

Contact Information
Samuel H. Payne, Ph.D.
samuel.payne@pnnl.gov

omics.pnl.gov