Spectral Alignment with Adaptive Penalties for Post-Translational Modifications and Mutations

Laurence E. Bernstein[1] and Nuno Bandeira[2]

[1] Bioinformatics Graduate Program, University of California, San Diego; Center for Computational Mass Spectrometry

[2] Department of Compter Science and Engineering, University of California, San Diego; Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego; Center for Computational Mass Spectrometry

Introduction

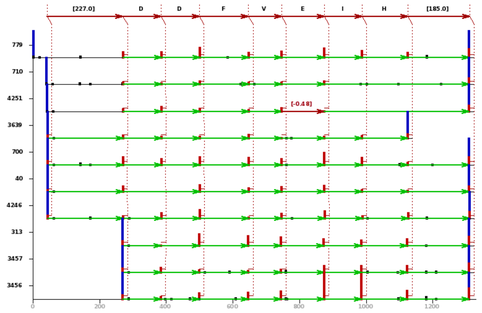
While high-throughput mass spectrometry is the leading method for protein identification, it remains challenging to process samples with unexpected post-translational modifications (PTMs) or distant homology to known proteins. As such, most search methods require pre-specified lists of allowed PTMs and a limit on the number of allowed PTMs per peptide. Unrestricted or "blind" search methods, while allowing any type of PTM still have restrictions on the number of modifications per peptide. We present an unrestricted, adaptive method for spectrum-sequence alignment that requires no a priori knowledge about the set of PTMs, and has no limitation on the location or number of modifications.

Methods

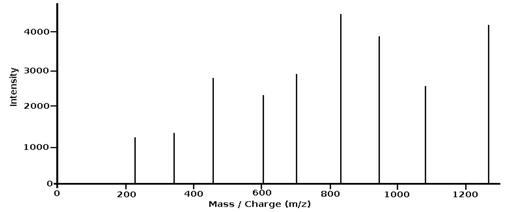
We present a new search approach combining multi-spectrum assembly, de novo sequencing and spectrum-sequence alignment with adaptive penalties based on sample-specific PTM and mutation frequencies. In addition, we incorporate tag-based filtering to reduce the search space and reduce false matches. Computational speed is significantly increased as alignments are only performed in the regions where the tags are matched. We use a dynamic programming approach for spectrum-sequence alignment where match scores are based on peak intensities combined with the sample-specific PTM and mutation penalties.

Multi-Spectrum Assembly

Spectra are aligned against each other in a pairwise fashion. All matching spectra are then assembled together and the aligned peaks are used to create a consensus spectrum. It is this spectrum which is used to perform alignment against the database.

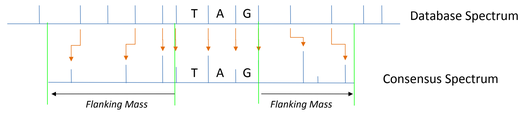


The resultant consensus spectrum is a far "cleaner" spectrum consisting only of those peaks for which matches were found in other spectra.



Tag Search

To speed up the database search we use tag-based filtering to decrease both the number of proteins to which the consensus spectrum is aligned as well as resticting the locations on the proteins that can form valid alignments.

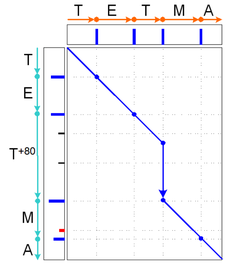


To perform the filtering we choose a tag length (L) and create all possible tags of that length from the consensus spectrum. A sequence of L+1 peaks whose peak mass differences are exact amino acid masses form a tag of length L. We find all proteins in the target database that contain a match to that tag and use the left flanking mass to determine a putative starting position for the alignment. During the alignment phase this starting location is allowed to vary by an amount that can be controlled by an external parameter.

Spectrum – Sequence Alignment

We align the consensus spectrum against the database sequence, by converting the sequence to a theoretical spectrum using the nominal amino acid masses. We then use a dynamic programming approach to find the best alignment between the two spectra with the score being adjusted positively for peak mass differences that match exactly while inexact matches (modifications) are penalized according to the size of the difference.

In the example alignment (to the right), an alignment with a single modification of +80 daltons is shown.



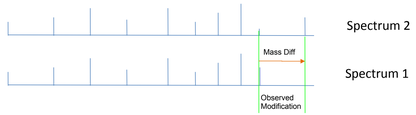
Modification Scoring Penalties

Modifications fall into three major categories: known, observed and unknown. Known modifications are modifications that are known to exist *a priori*. Although the alignment may be performed completely "blind" with no prior knowledge of modification sizes or their locations, prior knowledge of both modification size and associated amino acid(s) may be used if desired. Such modifications are scored with a much lesser penalty than other discovered modifications. Observed modifications are those that have some evidence of existence in the sample. Unknown modifications are those that do not fall into either the Known or Observed categories.

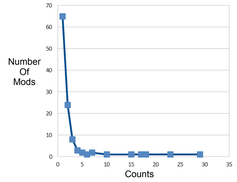
Modification Penalties	
Matching Peaks	+ + +
Known	-
Observed	- -
Unknown	- - -

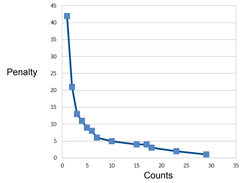
Observed Modification Penalties

The observed modifications are detected during the consensus spectrum assembly stage when two spectra match at one end but not at both. The difference in the total size of the two spectra is taken to be an Observed modification between the two spectra.



We convert the observed modifications to a modification distribution. From this data we can compute the probability that a given mass modification is likely to occur by calculating the probability that a mass modification with the same number of counts or less occurred in the training data. By using an inverse of these probabilities we can then create a penalty score. Since we can not know which of the two spectra has the modification we say that both a positive and negative mass modification of the same value have been observed.





"Gold Standard" for Results

In order to measure the accuracy of the consensus spectra, it is necessary to establish a "gold standard" for comparison purposes. To do this we first run the MODa software at 1% PSM level FDR and then manually curate the post-translational modifications from the MODa frequency table. The selected modifications are then used as input to the MS-GFDB software and MS-GFDB is run at 1% FDR to establish a "truth" set for the individual spectrum. To compare our consensus spectrum results to this truth set we match the spectra that comprise our consensus spectra to the MS-GFDB results to see if our consensus spectra map to the same protein and location as the MS-GFDB result.

Data Set	Curated Modifications				Number of Spectra Found at 1% FDR
LENS	-43	C	32	W	1303
	-18	ST	42	(N-term)	
	-17	Q (N-term)	43	(N-term)	
	1	NQ	44	W	
	14	KR	80	STY	
	16	MW			
aBTLA	-91	C	22	*	270
	-17	Q (N-term)	28	KRN	
	1	NQ	42	(N-term)	
	14	K	43	(N-term)	
	16	MW			

Tag Filtering Results

Tag Length	Tag Gaps	Flank Match	Tags Found LENS	Contigs Correct LENS	% Correct LENS	Tags Found aBTLA	Contigs Correct aBTLA	% Correct aBTLA
3	0	0	2532	241	91.985	2958	114	96.491
3	0	1	474	200	78.125	1002	93	92.079
4	0	0	335	242	94.628	606	102	98.077
4	0	1	159	187	95.897	424	84	92.308
4	1	0	1911	251	92.279	2316	111	94.058
4	1	1	580	209	78.868	968	93	87.736
5	0	0	133	204	97.608	311	83	100.00
5	0	1	102	176	96.175	260	66	100.00
5	**1**	**0**	**335**	**242**	**96.414**	**616**	**106**	**97.248**
5	1	1	221	200	89.286	469	89	92.708

We experimented with various values for tag length, allowable numbers of gaps in the tag, and flanking mass matches (flanking mass from the consensus spectra matching the amino acid sequence mass in the database). Longer tag lengths without gaps yielded the highest accuracy, but with reduced numbers of total contig matches, whereas allowing tag gaps led to more matches but lower accuracy. An excellent compromise was found by using a longer (5) tag length, with one gap allowed.

Alignment Results

Modification	Frequency	Explanation
18	5	Reversal Artifact *
-2	3	N/A **
14	2	Methylation
16	2	Oxidation
42	2	Acetylation

* Artifact of improper spectrum reversal during assembly
** Artifact of Low Mass Accuracy Data

Conclusions

We demonstrate a blind search algorithm that combines multi-spectrum assembly, tag-filtering and spectrum-sequence alignment to perform database matching of highly modified spectra. We show that the parameters of tag filtering for assembly-derived contig consensus spectra are most effective when set to ignore flanking masses while still allowing for missing peaks in the tags. Tags of longer length still perform very well because the consensus spectra eliminate the deleterious effect of spectrum-level modifications. This has a significant impact on both the overall speed of the algorithm (in terms of number of alignments that must be performed) and the accuracy of the alignment. The consensus alignments find many of the modifications that we expect in the sample, although we also detect two artifacts (reversal and low-mass spurious offsets) that will be addressed in later work.

Dataset Information

LENS: 17161 spectra from a 93 year old subject analyzed using two-dimensional liquid chromatography, and collected on both high and low mass accuracy instruments. [1]

ABTLA: 27640 proteins comprising the light chain from human antibodies analyzed using an LTQ XL-Orbitrap mass spectrometer.

[1] Searle, B.S. et al. Identification of protein modifications using MS/MS de novo sequencing and the Opensea alignment algorithm. J. Proteome Res. 4, 546-554 (2005).

[2] Bandeira, N. et al. Automated *de novo* protein sequencing of monoclonal antibodies. Nature Methods. Vol 6, Num 12, 1336-1338 (2008).