Flexible Library Assisted SearcH (FLASH): a hybrid library/database search engine

Joon-Yong Lee, Grant M Fujimoto, Christopher S Wilkins, Richard D Smith, Samuel H Payne
Pacific Northwest National Laboratory, Richland WA

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Overview

- Hybrid library/database search for bottom-up proteomics
- Combine sensitivity of spectral library search with the statistical rigor of peptide database search
- Optimized algorithm runs 5x faster than database search engine
- Improved sensitivity and specificity when searching very large libraries/databases

Introduction

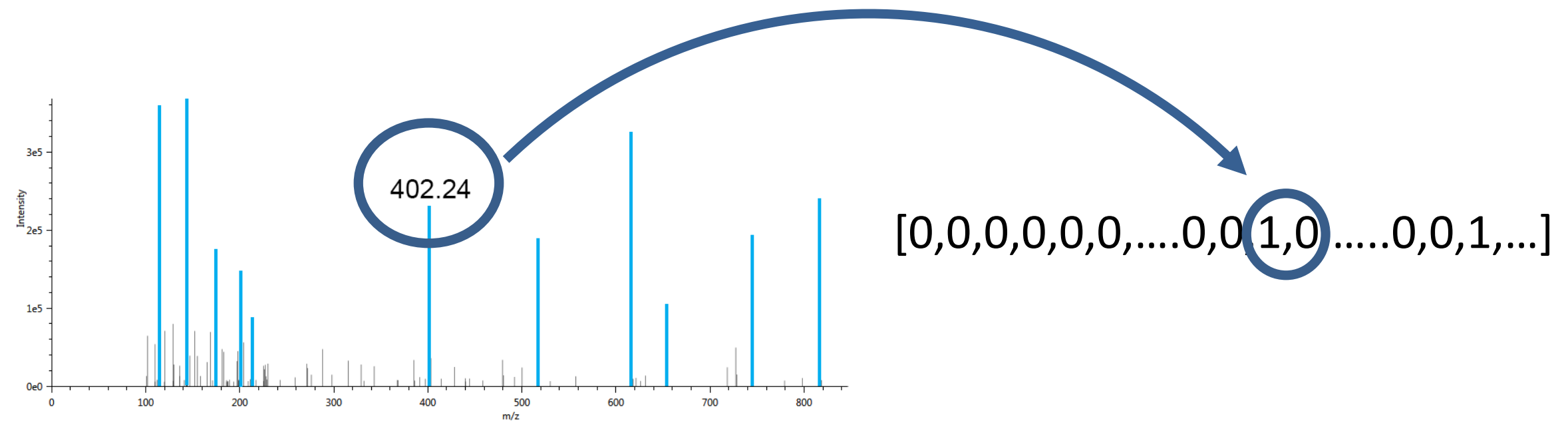
- Public proteomics data is increasing peptide library coverage for proteins across many organisms.
- Library and database search algorithms have historically been distinct, each with unique benefits.
- Library search algorithms are more sensitive than database search algorithms.
- Library searches lack a rigorous statistical probability method.

Methods

The FLASH algorithm is a hybrid of library and database search, meaning it uses both spectrum/spectrum matching and traditional database scoring. The library search portion of the algorithm rapidly identifies candidate peptide annotations for query spectra by finding similar spectra in the library. The database search portion uses the statistically rigorous generating function of MSGF+ to evaluate the proposed peptide annotation.

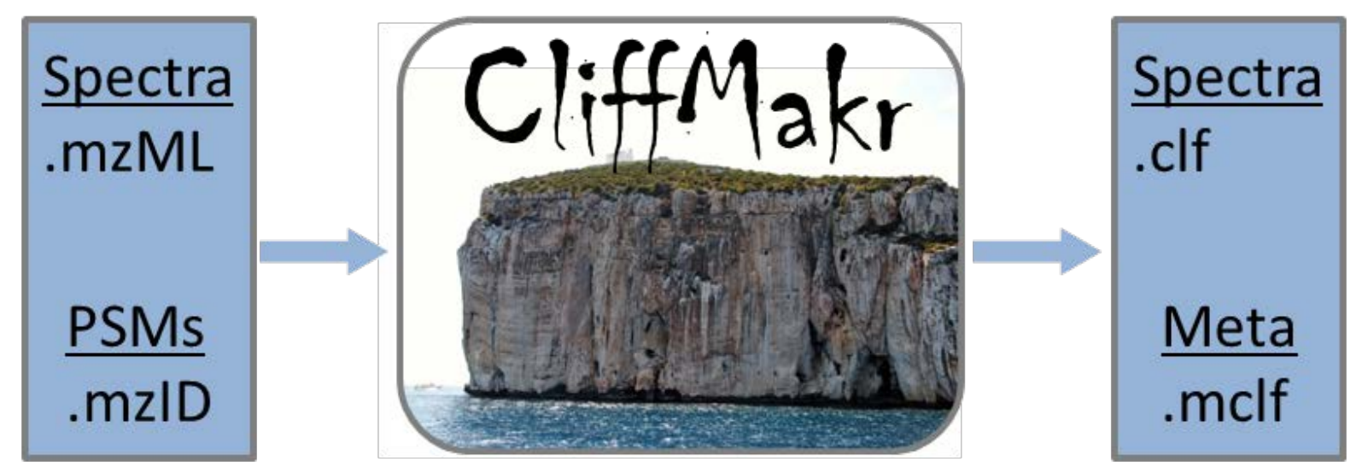
Binary simplification of spectra

- Discretize spectra to 0.05 mz bins
- Record top 20 peaks as binary



The Condensed Library Format

- Input trusted PSMs with .mzID, .mzML
- Binarize spectra in .clf
- Keep meta-data in .mclf



Hybrid library and database search engine (FLASH)

- Spectrum/spectrum matching. The Blazing Signature Filter (BSF) counts shared peaks between all query and library spectra (regardless of precursor m/z differences).
- Infer annotations. For significant query/library pairs, we calculate the Δmass. If the Δmass is close to zero, the library annotation is considered for the query spectrum. Otherwise, we apply Δmass to the library annotation.
- All candidate PSMs of query spectra are tested with the Generating Function from MSGF+ to obtain a statistically rigorous spectrum E-value, output in .mzID format.

Results

We created FLASH, which combines the benefits of both a library and a database search algorithm. FLASH is able to identify the correct peptide/spectrum match rapidly and in the presence of an overwhelming background of unrelated peptides.

Shared peaks as an efficient filter

The first step of FLASH is to compare query spectra against the annotated library. To perform spectrum/spectrum comparisons fast enough to keep pace with growing library sizes, we simply measure spectrum/spectrum similarity as shared peak count using the Blazing Signature Filter (below).

Spectrum/spectrum matches of the same peptide are easily distinguished from random spectral pairs. However, we want to use FLASH to also detect spectrum/spectrum matches of similar peptides, e.g. PTMs or amino acid substitutions.

Spectra from similar peptides will share many peaks and are commonly identified via spectral alignment (Figure 1).

A low mass cutoff for fragment peaks improves specificity in determining similar spectra (1-2 amino acid mismatches) from random peaks (Figure 2).

Blazing Signature Filter (BSF)

To efficiently compute shared peak count, we use the Blazing Signature Filter – an optimized set overlap algorithm. Using the BSF allows us to identify both exact matches and similar spectra from related peptides (non-exact matches e.g. PTMs).

After conversion of spectra into binary arrays (see Methods), the BSF rapidly identifies shared peaks using bit-wise operators (Figure 3).

Query/Library pairs that pass a shared peak threshold are further evaluated by using MSGF+ to score the peptide annotation and obtain a rigorous e-value score.

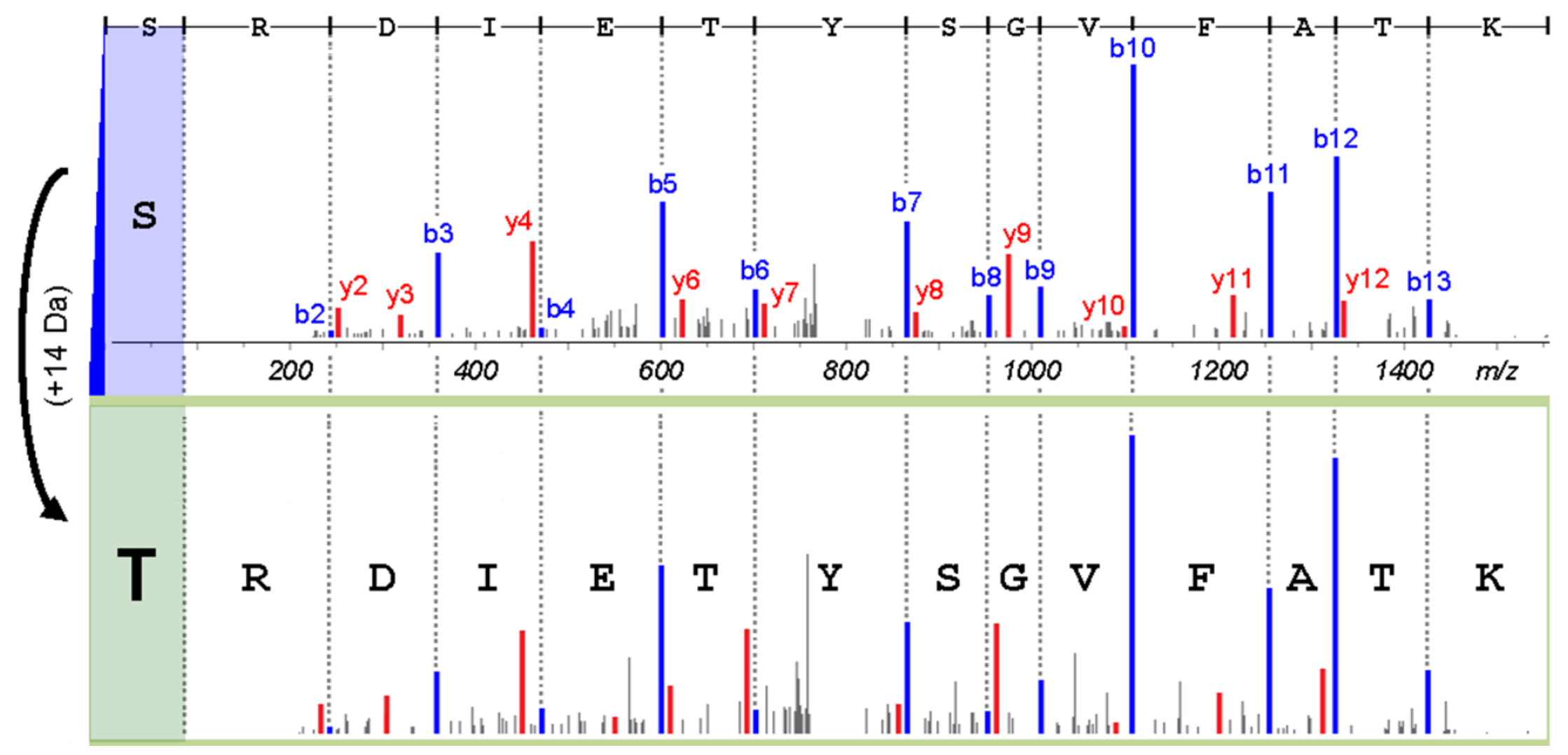


Figure 1 – Spectral alignment of similar PSMs. Amino acid substitutions and PTMs correspond to a mass difference, which is observed an m/z shift in half of the fragment peaks. The other half of the peaks align trivially with the reference spectrum.

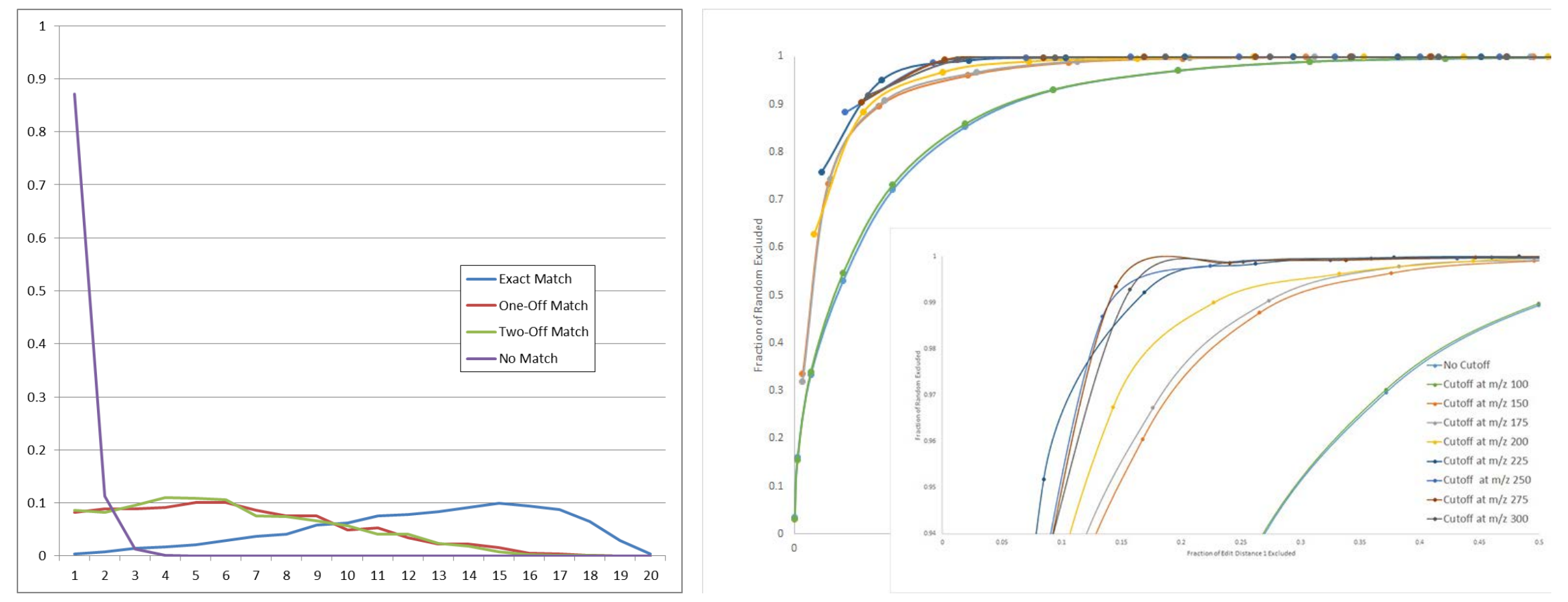
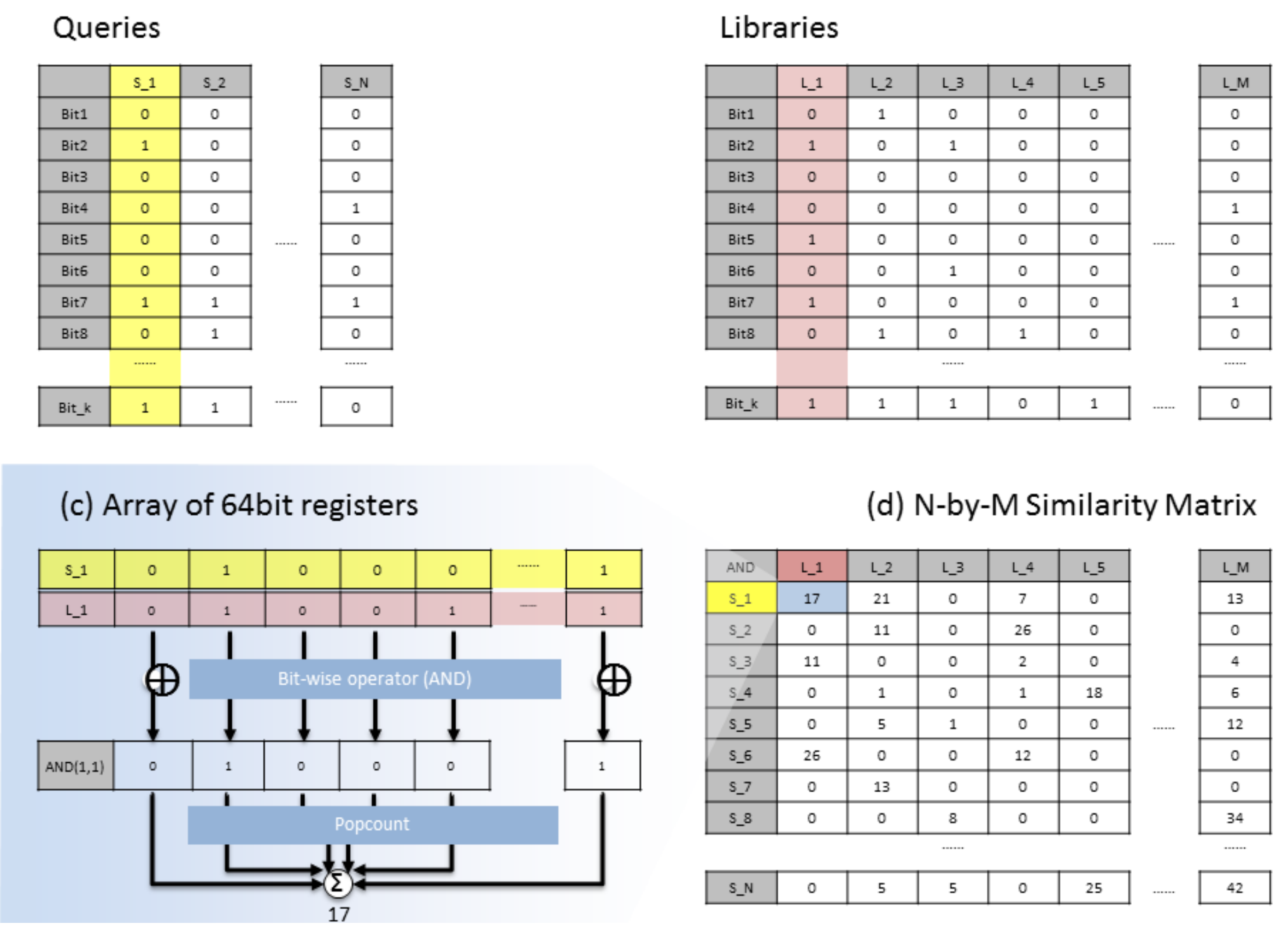


Figure 2 – Spectral similarity of related peptides. Related peptides have a higher shared peak count than random (left). A low mass cutoff removes non-specific peaks and improves accuracy (right).



Queries

	S_1	S_2		S_N
Bit1	0	0		0
Bit2	1	0		0
Bit3	0	0		0
Bit4	0	0		1
Bit5	0	0	......	0
Bit6	0	0		0
Bit7	1	1		1
Bit8	0	1		0
	......			......
Bit_k	1	1	......	0

Libraries

	L_1	L_2	L_3	L_4	L_5		L_M
Bit1	0	1	0	0	0		0
Bit2	1	0	1	0	0		0
Bit3	0	0	0	0	0		0
Bit4	0	0	0	0	0		1
Bit5	1	0	0	0	0	......	0
Bit6	0	0	1	0	0		0
Bit7	1	0	0	0	0		1
Bit8	0	1	0	1	0		0
			......				......
Bit_k	1	1	1	0	1	......	0

(c) Array of 64bit registers

S_1	0	1	0	0	0	......	1
L_1	0	1	0	0	1	----	1
AND(1,1)	0	1	0	0	0		1

(d) N-by-M Similarity Matrix

AND	L_1	L_2	L_3	L_4	L_5		L_M
S_1	17	21	0	7	0		13
S_2	0	11	0	26	0		0
S_3	11	0	0	2	0		4
S_4	0	1	0	1	18		6
S_5	0	5	1	0	0	......	12
S_6	26	0	0	12	0		0
S_7	0	13	0	0	0		0
S_8	0	0	8	0	0		34
			......				......
S_N	0	5	5	0	25	......	42

Figure 3 – Binary set comparison. The binarized spectra from the query and library are efficiently compared to count shared peaks.

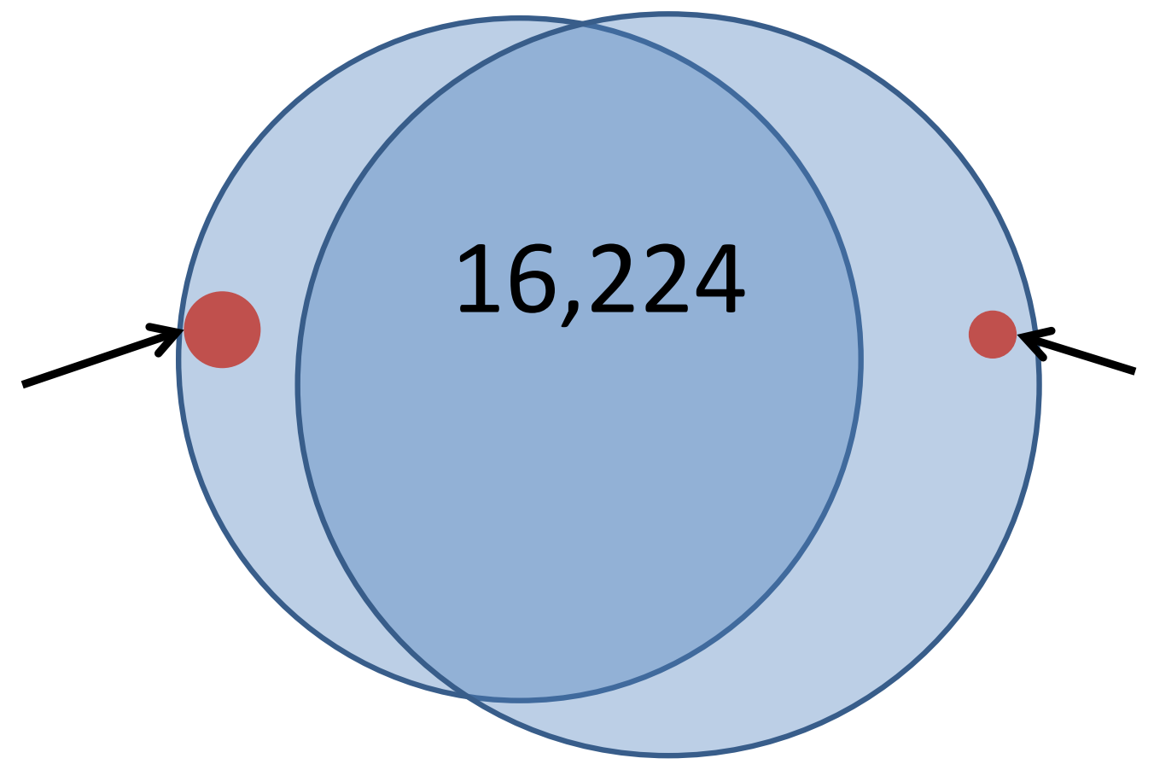
Sensitivity and specificity with large libraries

To challenge the ability of FLASH to correctly identify peptide/spectrum matches in the presence of large and diverse background noise, we created a library with >1,000,000 peptides from 48 bacterial species. We compared the performance of FLASH with MSGF+ using a database that comprises the proteome of the 48 bacteria.

B. Cereus searched versus Biodiversity library
~42,000 MS/MS spectra, Thermo QExactive

Run time: 20 hrs (MSGF+) 4 hrs (FLASH)

MSGF+ only
1,746 PSMs
123 false-positives
FDR ~ 0.6%



FLASH only
4,330 PSMs
52 false-positives
FDR ~ 0.2%

As an additional test, we searched spectra from H. sapiens against the bacterial biodiversity library. Consistent with the specificity observed above, only 24 out of 42,000 MS/MS had a significant MSGF+ score. The peptide annotation for these spectra come from metabolic enzymes, whose sub-sequence is conserved from bacteria to humans.

Conclusions

- FLASH merges the best attributes of library and database search algorithms
- Rapid and specific spectrum annotation with very large sequence search space
- Libraries can be created from any search engine (.mzID, .mzML)

Acknowledgements

This project was supported by the U.S. Department of Energy (DOE) Office of Biological and Environmental Research, Early Career Research Program and the Pan-omics program at PNNL. PNNL is a multiprogram national laboratory operated by Battelle for the DOE.

Career Opportunities: visit http://omics.pnl.gov/careers

CONTACT: **Samuel H Payne, Ph.D.**
Biological Sciences Division
Pacific Northwest National Laboratory
E-mail: samuel.payne@pnnl.gov

http://omics.pnl.gov
@omicsPNNL