Implementation of a Charge-State and Sequence-Dependent Scoring System Dramatically Improves Matching of Peptide ETD Spectra

Robert J. Chalkley, Peter R. Baker, Katalin F. Medzihradszky, and A. L. Burlingame

UCSF

Introduction

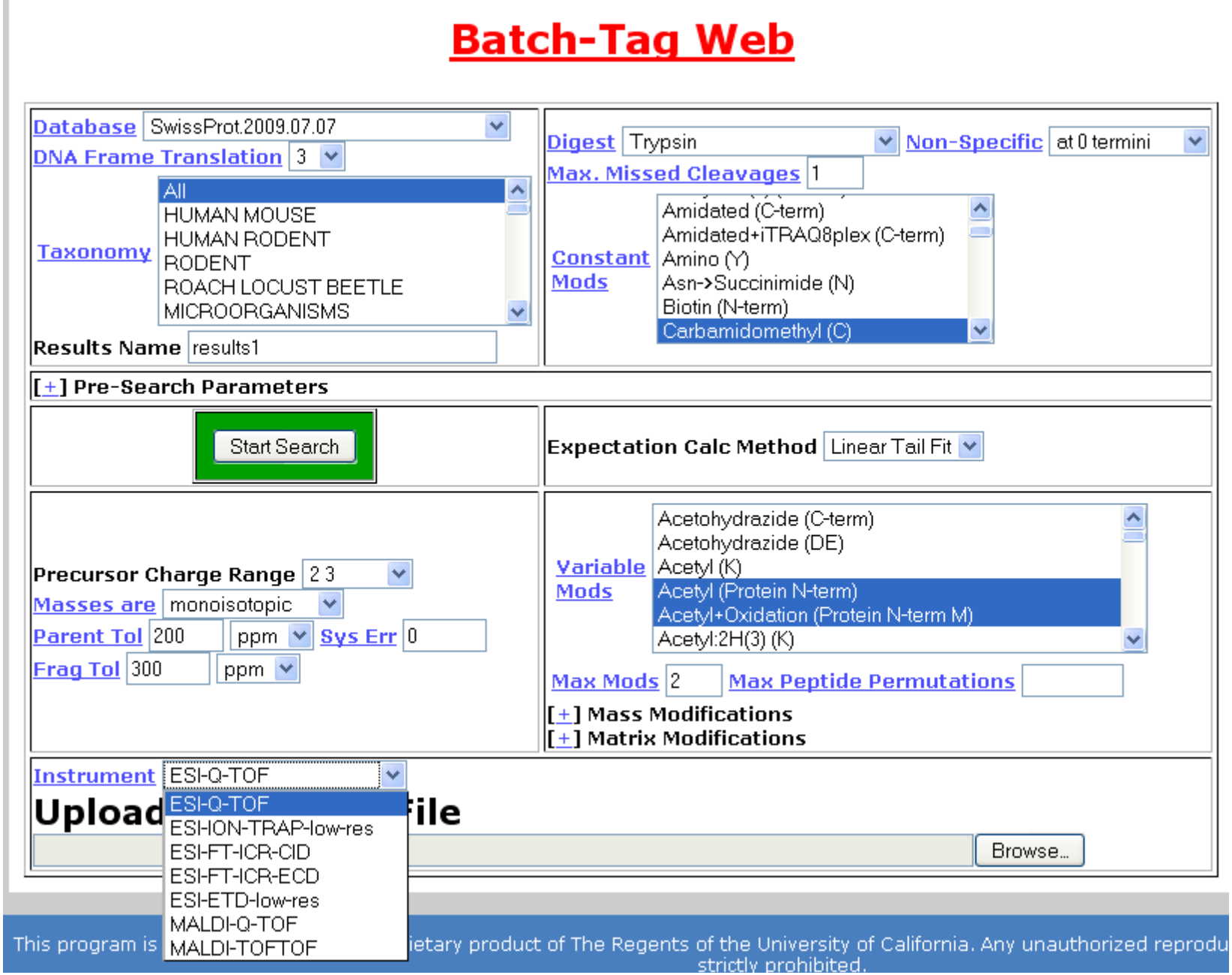
- Software for the analysis of CID data is well developed
  - Several studies published on fragment ion frequencies, sequence-dependent fragmentation, effect of charge state (mobile protons)…
- ETD and ECD are becoming increasingly used
  - Complementary results
  - Effective for peptides bearing labile PTMs
  - Effective for large peptides / small proteins
- Software for ETD data analysis is still being developed / optimized

Improving Software for Peptide ETD/ECD Analysis

1. Characterize the frequency of observation of different fragment ion types in ETD data from different types of peptides

2. Use this information to improve ion weighting / scoring in Protein Prospector software for analysis of ETD/ECD data

3. Evaluate performance of new scoring compared to previous version and compared to other search engines on common dataset.



•Different ion types and scoring considered depending on the instrument type

Scoring in Batch-Tag

- Matching of each ion type accrues a different score.
- Ion match scores are summed together to give a peptide score.

- Spectrum is also searched against a decoy (sequence-randomized) database
- Results of all scores against decoy database are fit to a distribution

- Probability (and then expectation value) calculated for match of a given score to be part of the random distribution.

Chalkey et al. *Mol Cell Proteomics* (2008) **7** 12 2386-2398

ETD Fragmentation Training Data

- Mouse Synaptic and Nuclear Preparations digested with four different cleavage specificities:
  - Tryptic
  - Lys-C
  - Lys-N
  - CNBr

- Data acquired on LTQ-Orbitrap, with precursor ions measured in Orbitrap and fragments measured in LTQ (supplemental activation on).
- FDR rates calculated using database with random sequences concatenated onto normal database.

Chalkley et al. *Anal Chem* (2010) **82** 2 579-584

Frequency of Occurrence of Different Ion Types in ETD Data

Enzyme	LysC	LysC	LysN	LysN	CNBr	CNBr	Trypsin	Trypsin
Prec. Charge	2+	>2+	2+	>2+	2+	>2+	2+	>2+
b	0.02	0.03	0.06	0.04	0.04	0.05	0.03	0.03
c-1	0.11	0.05	0.17	0.06	0.14	0.08	0.10	0.04
c	0.20	0.33	0.49	0.42	0.31	0.36	0.11	0.34
y	0.13	0.09	0.05	0.06	0.06	0.06	0.14	0.11
z	0.30	0.29	0.17	0.27	0.27	0.25	0.31	0.31
z+1	0.23	0.20	0.07	0.15	0.18	0.21	0.32	0.17

- Basic residue location (enzyme) significantly affects ion frequency
- Hydrogen transfer products more common in spectra of 2+ precursors

Chalkey et al. *Anal Chem* (2010) **82** 2 579-584

Implementing New Scoring System

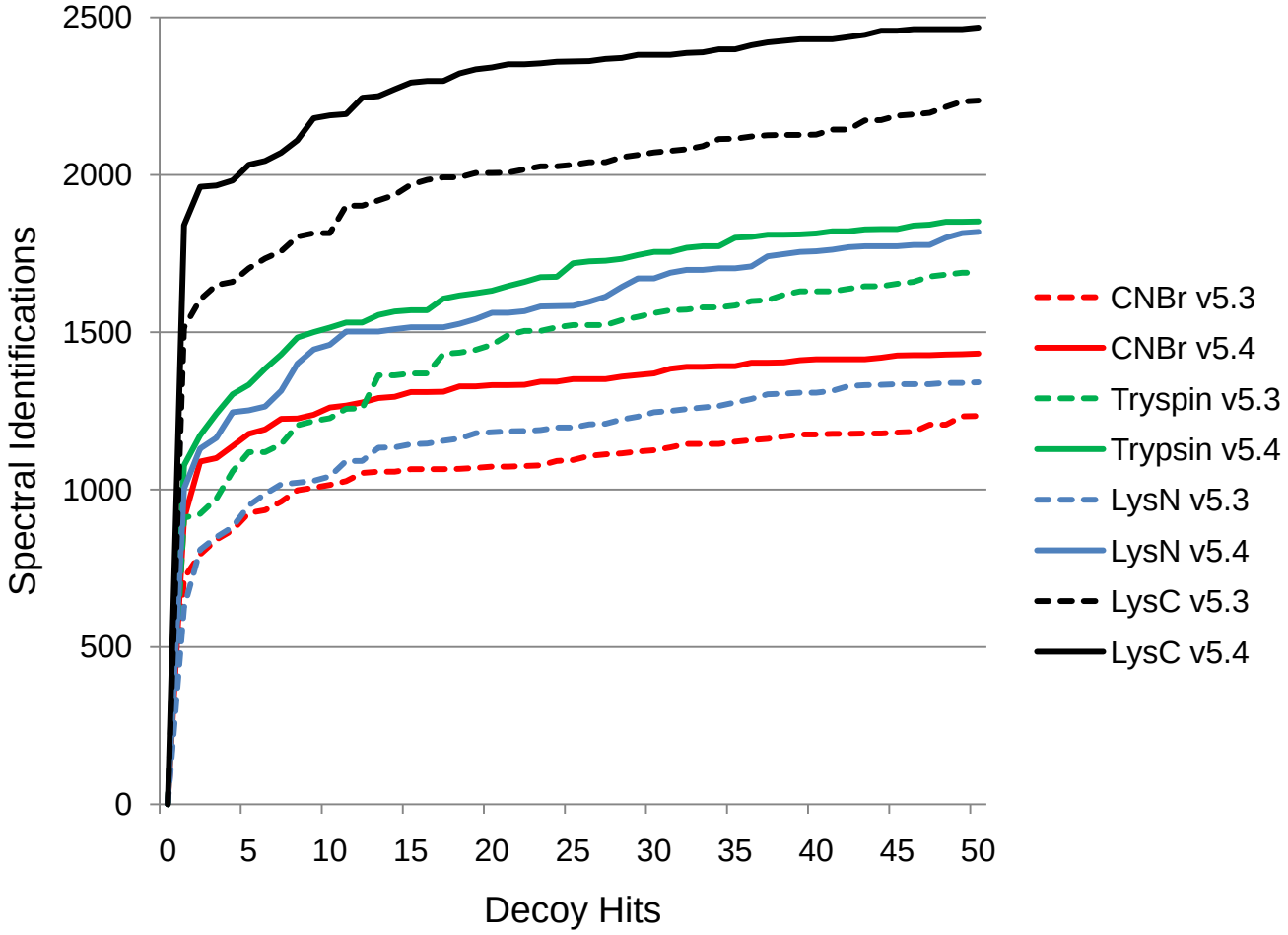
•Fragment ion frequencies used to create new scoring (PPv5.4) which uses different weighting depending on precursor charge-state and presence of basic residues at termini.

Precursor Charge	Basic Residue at:
2	C-terminus
3	N-terminus
4	N- and C-termini
5+	Neither terminus

•Results compared to using scoring system based on average ion frequencies in tryptic ETD spectra (PPv5.3)

- •No differential scoring based on sequence and charge
- •Bias toward peptides with basic C-terminal residue

Effect of Implementing New Scoring System



- Improvements seen for peptides produced by all cleavages
- Dramatic improvement for Lys-N peptides: 39% more IDs at 1% FDR

Comparison of Results to Other Search Engines

Previously published comparison of four search engines:

- Mascot
- OMSSA
- Spectrum Mill
- X!Tandem

- Combination of several datasets
  - tryptic and Lys-C peptides
  - phosphopeptide-enriched samples.

  - ETD data acquired on Agilent 6340 3D ion trap (with supplemental activation)
  - FDR estimates calculated using separate normal and decoy database searches, then combining for setting thresholds.

[1]Kandasamy et al. *Anal Chem* (2009) **81** 17 7170-7180

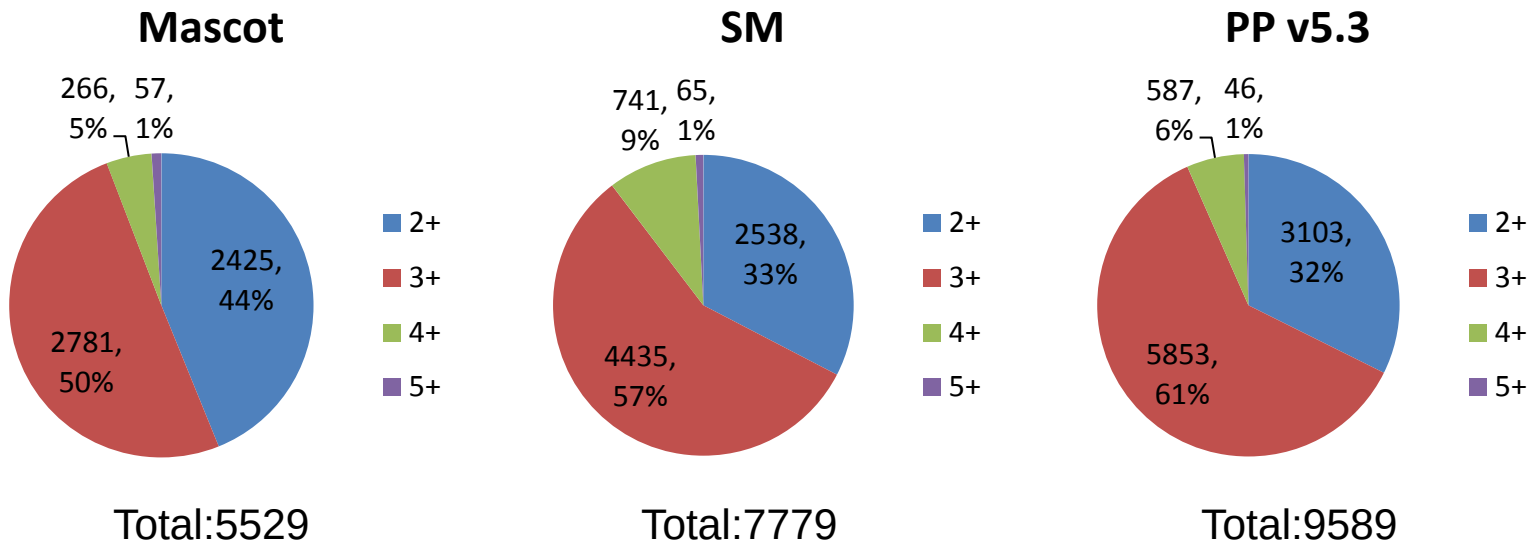
Comparison to Other Search Engines

	OMSSA	Mascot	Spectrum Mill	X!Tandem	PP v5.3	PPv5.4
All Data, 1% FDR threshold	3.7 E 10-3	41.2	10.1, 10.3, 12.5, 12.5	-2.797	0.075	0.028
Total spectral IDs at 1% FDR threshold	4491	5529	7779	4997	9653	14095

- Both versions of Protein Prospector outperformed other search engines
- New scoring increased number of identifications by over 40% at 1% FDR threshold.

Values for search engines other than Protein Prospector were derived from Kandasamy et al. *Anal Chem* (2009) **81** 17 7170-7180

Breakdown of Spectral Identifications by Charge-State



(Values for search engines other than Protein Prospector are derived from results from Kandasamy et al. *Anal Chem* (2009) **81** 17 7170-7180)

Overlap in Results Between Different Search Engines

At 1% FDR:

No. of Search Engines Id'd Spectrum	Total	Search Engine Combination	No. Spectra	Search Engine Combination	No. Spectra	Search Engine Combination	No. Spectra	Search Engine Combination	No. Spectra
all 5	1815								
Only 4	1489	O, P, S, X M,O,S,X	676 16	M,P,S,X	469	M,O,P,S	246	M,O,P,X	82
Only 3	2320	M,P,S M,P,X O,S,X	924 105 13	P,S,X M,O,P M,O,X	559 48 2	O,P,S M,O,S	409 32	O,P,X M,S,X	211 17
Only 2	3976	P,S M,S O,X	1344 122 16	M,P S,X M,X	1195 36 5	O,P O,S	613 36	P,X M,O	585 24
Only 1	6385	P O	4724 205	S	790	M	427	X	239

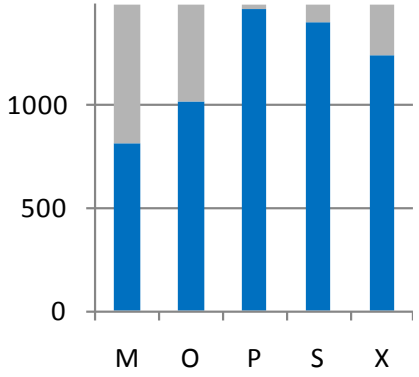
Total: 15985

M=Mascot; O=OMSSA; P=Prospector; S=Spectrum Mill; X=X!Tandem

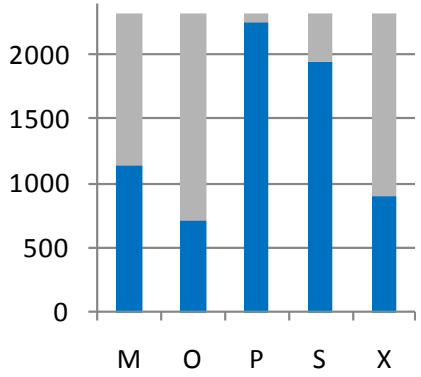
- About 11% identified by all search engines
- Over 90% of the IDs identified by only 2 search engines were Protein Prospector + one other search engine.

Search Engine Membership in Different Levels of Agreement

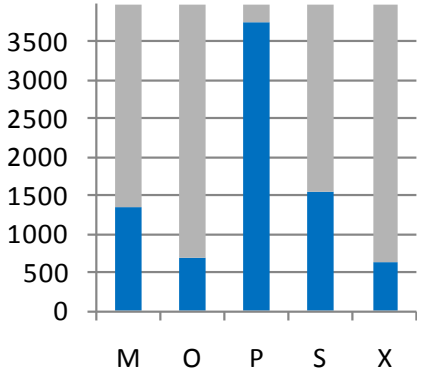
Four Agree



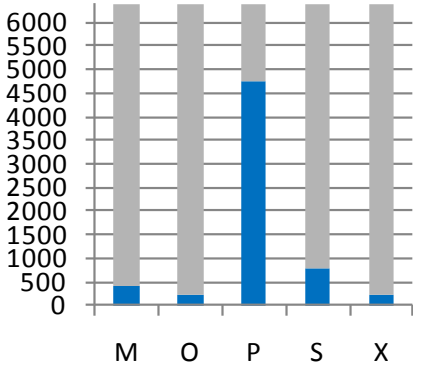
Three Agree



Two Agree



Unique



- Very few identifications by more than one search engine did not include Protein Prospector.
- Mascot is the most common search engine to miss out on an identification made by all other search engines
  - Is there a spectrum type that it performs badly on?

Conclusions

- ETD fragmentation of doubly-charged precursors is very sequence and charge-state dependent.
- Fragments observed from 2+ precursors differ significantly from those from a higher charge-state.

- Implementation of charge- and sequence-dependent scoring causes a noticeable improvement in search engine performance for all data analyzed.

- Protein Prospector appears to dramatically outperform alternative software options (last year's versions).
    - Many of the identifications were found by Protein Prospector were found by at least one other search engine.

Future Directions

- Apply similar analysis to CID data
  - Sequence and charge-state dependent CID scoring
- Integrating CID and ETD Results
  - How to combine CID and ETD assignments to the same precursor?

Acknowledgements

We thank Ralf Schoepfer and Sam Myers for samples used in the enzyme comparison dataset and Henrik Molina for supplying us with his search results from his published search engine comparison.

This work was funded by NIH NCCR grant P41 RR01614

Vincent J. Coates Foundation.

This software is freely available at http://prospector.ucsf.edu