

# Roles for sequence tagging in protein identification

[david.i.tabb@vanderbilt.edu](mailto:david.i.tabb@vanderbilt.edu)

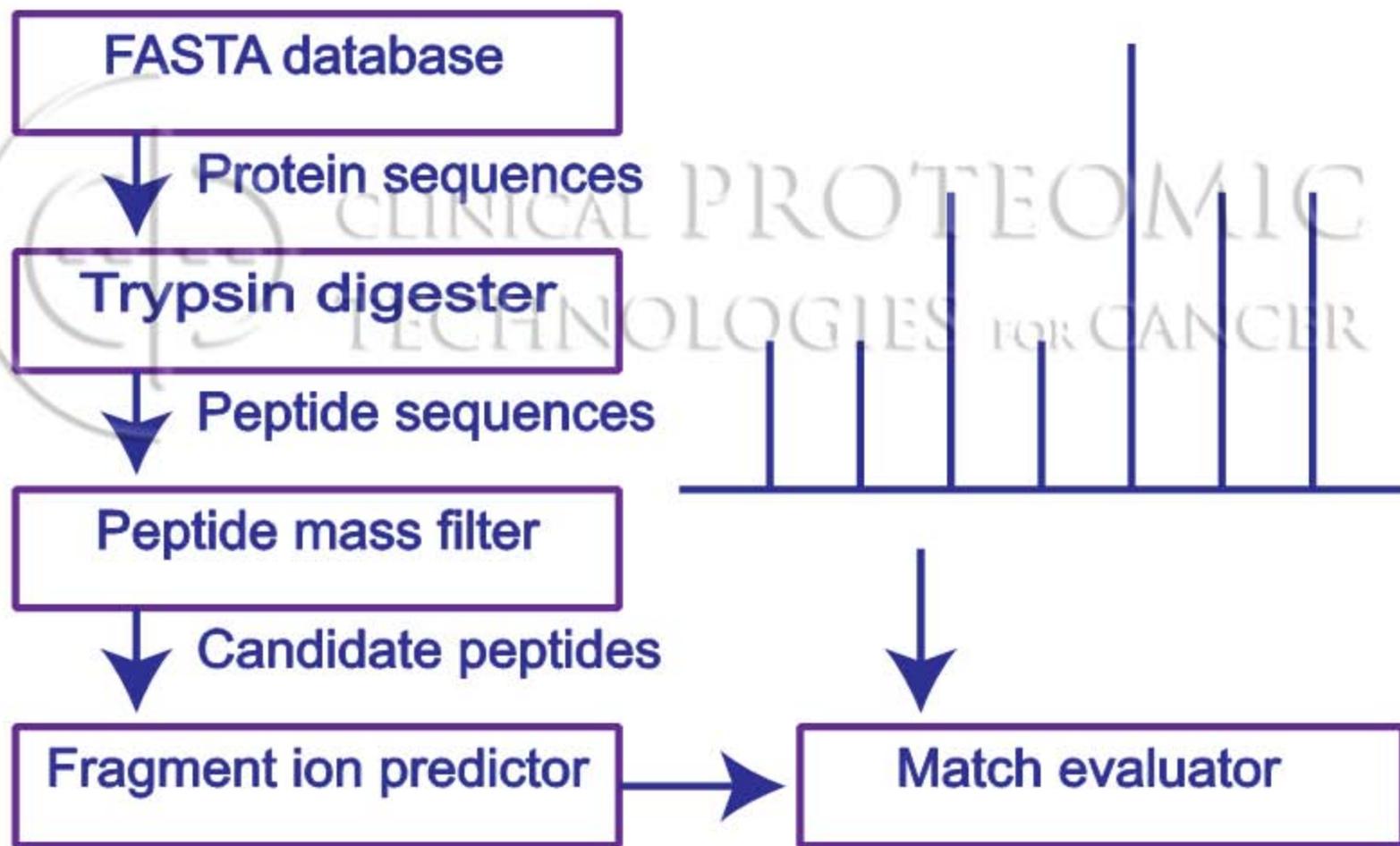
Biomedical Informatics

Mass Spectrometry Research Center

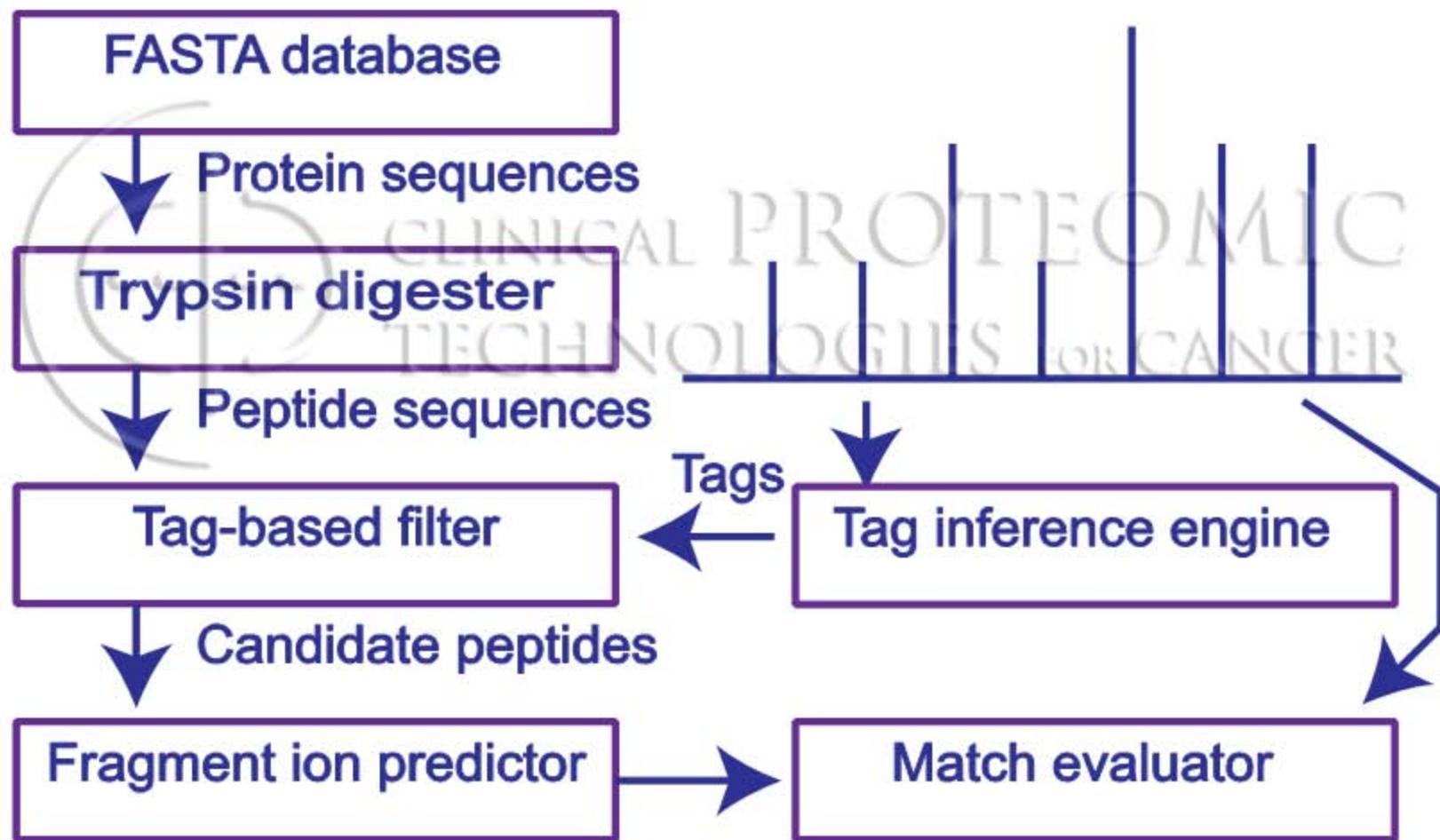
# Overview

- Review of sequence inference algorithms
  - Spectrum-graph systems
  - Direct scoring systems
- Applications for sequence tagging
  - Quality filtering
  - PTM hunting
  - Mutation hunting

# Database search overview



# Sequence tagging overview

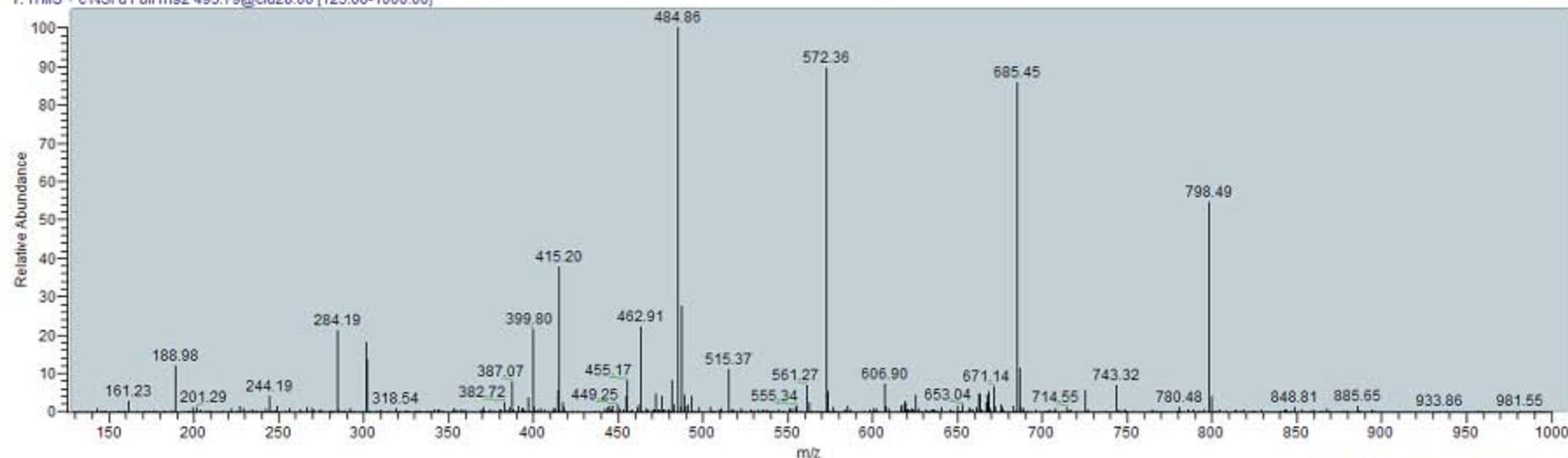


# Tags are partial sequences

- Interpeak  $m/z$  differences are residue masses. Tag includes flanking masses.
- But how do we infer these?

G | L | L

Mc\_CPTAC\_062407o\_final\_run3#6072 RT: 58.00 AV: 1 NL: 7.12E3  
T: ITMS + c NSI d Full ms2 493.79@cid28.00 [125.00-1000.00]



# Spectrum graph-based inference

- C Bartels, *Biomed. Environ. Mass Spectrom.* (1990) 19:363-368.
- Applies machine learning strategies to learn how to use information of spectrum.
- Combines information among related fragment ions to evaluate probability of breakage at given mass within peptide.
- Builds a graph summarizing likely peptide breakpoints and reads it to list sequences.

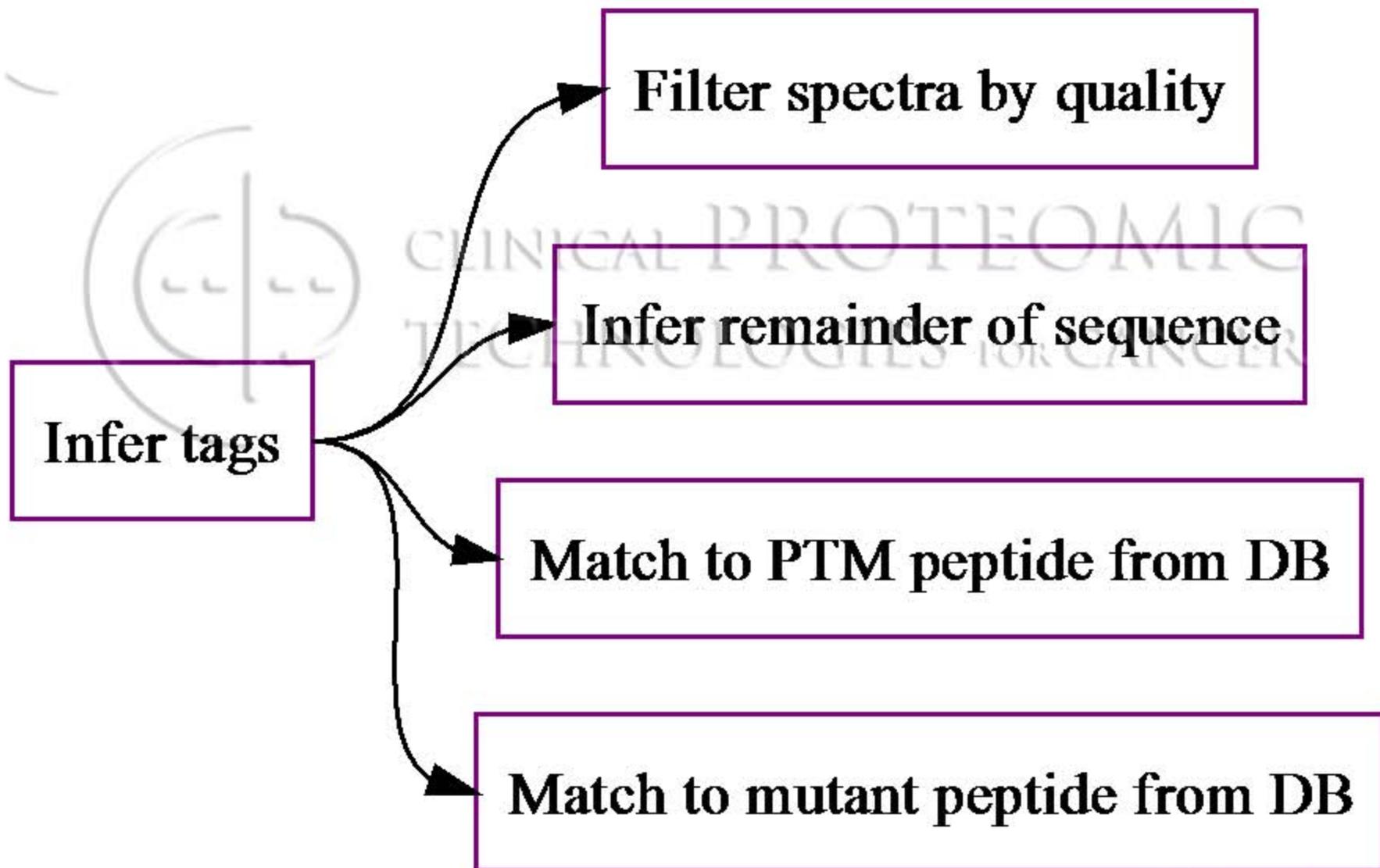
# Tools employing spectrum graphs

- Lutefisk: JA Taylor, *Rapid Comm. Mass Spectrom.* (1997) 11: 1067-1075.
- PepNovo: A Frank, *Anal. Chem.* (2005) 77: 964-973.
- InsPecT: S Tanner, *Anal. Chem.* (2005) 77: 4626-4639.
- NovoHMM: B Fischer, *Anal. Chem.* (2005) 77: 7265-7273.

# Direct scoring inference tools

- *Sequences are scored directly against observed peaks, not a graph abstraction.*
- Peaks: B Ma, *Rapid Comm. Mass Spectrom.* (2003) 17: 2337-2342.
- GutenTag: D Tabb, *Anal. Chem.* (2003) 75: 6415-6421.

# I've inferred tags. Now what?



# Quality filtering

- Filters out spectra that lack reliable tags
- Reduces spectral collections prior to database search, speeding processing
- Enables subsequent examination of unidentified spectra with high tag scores
  
- Employed in Spectrum Mill

# Tag vs. database reconciliation

- Peptide does not contain tag sequence: **STOP**
- Peptide contains tag sequence:
  - Both terminus masses match: **Match as-is**
  - Neither terminus mass matches: **STOP**
  - Only N-terminus mass matches: **Modify C**
  - Only C-terminus mass matches: **Modify N**

# PTM reconciliation

- DB sequence: QVLEVTALVVK
- Tag match: [QVLE-17]VTALVVK
- The tag **VTA** matches this database sequence, and the C-terminus corresponds exactly in mass.
- Testing different possible locations for 17 loss may reconcile N-terminus as a pyroglutamine.

# Mutation reconciliation

DLFSNA**A**IDE INEK from the DB contains tag “**AID**”, but the mass observed to the tag N-terminus is too high by 14 Da. Test these sequences:

DLFS**Q**AIDE INEK

DLF**T**NAIDE INEK

**E**LFSNAIDE INEK

# Evaluate the full-spectrum match

- Ordinary peptides from the database compete with mutated sequences.
- For the next challenge, how do we differentiate these matches?
- **AEMADQAS\*AWLTR** (\* = Ser to Ala mutant)
- **AEMADQAAAWLTR**

# What's missing?

- Difficult to integrate tag-based identifications with DB search results
- Unclear whether FDR assessment must be modified in tag-based interpretation
- Challenging to improve sequence inference accuracy
- Necessary to convince proteomics community of gains through tagging

# Summary

- Sequence tagging may greatly expand the scope of peptides we can identify.
- Tagging will not replace DB search, but it can complement it powerfully.
- If a fraction of the effort sunk into DB search scoring were in tagging instead, we would know more about our data!

# Acknowledgment

- NIH/NCI 1 R01 CA126218
- NIH/NCI 1 U24 CA126479
- NIH/NIEHS P30 ES000267
- American Cancer Society Institutional Research Grant (#IRG-58-009-48)
- Matthew Chambers and Zeqiang Ma