CMSC423

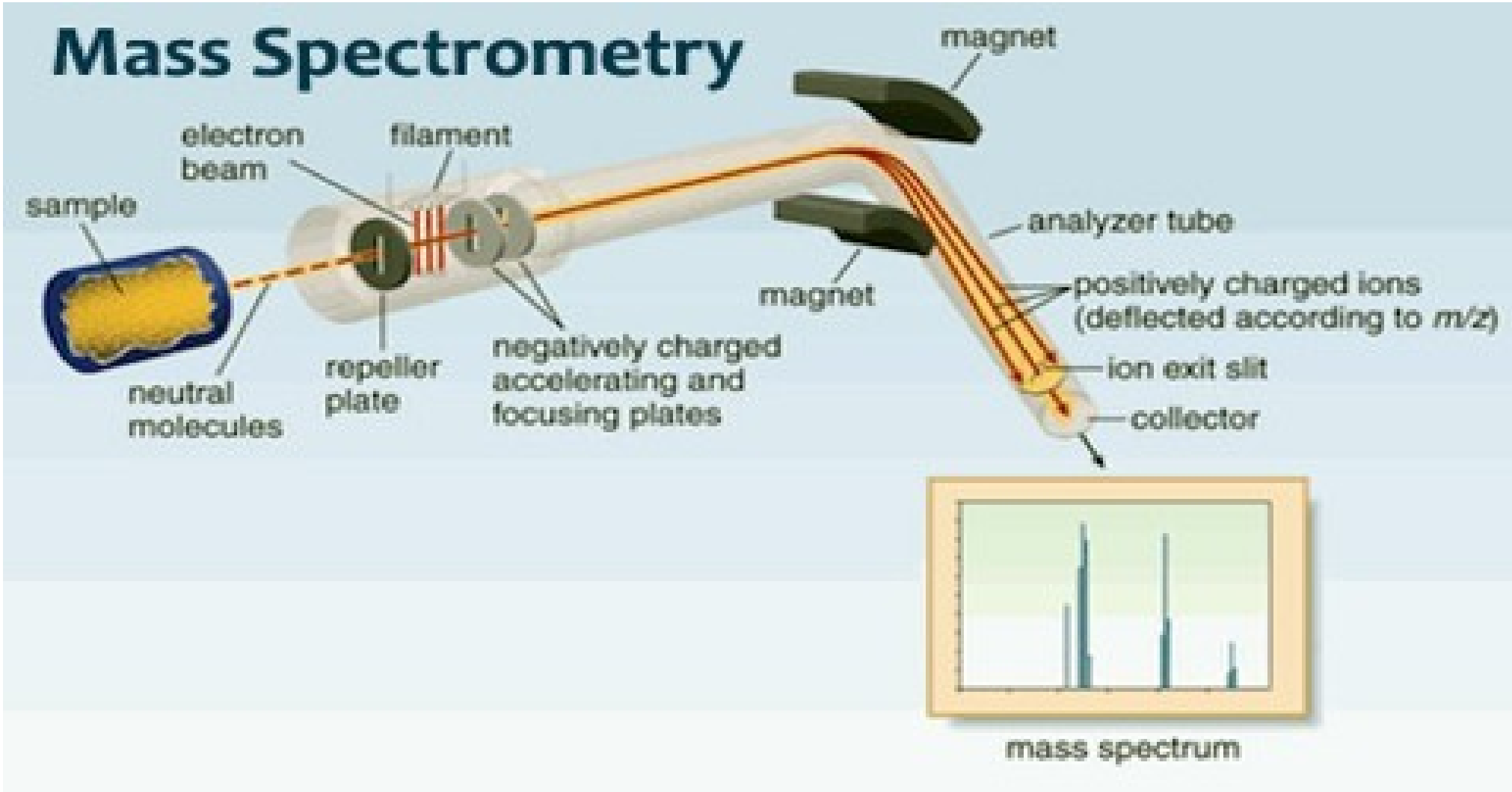
Chapter 4 – Proteomics/mass-spectrometry Leaderboard searching

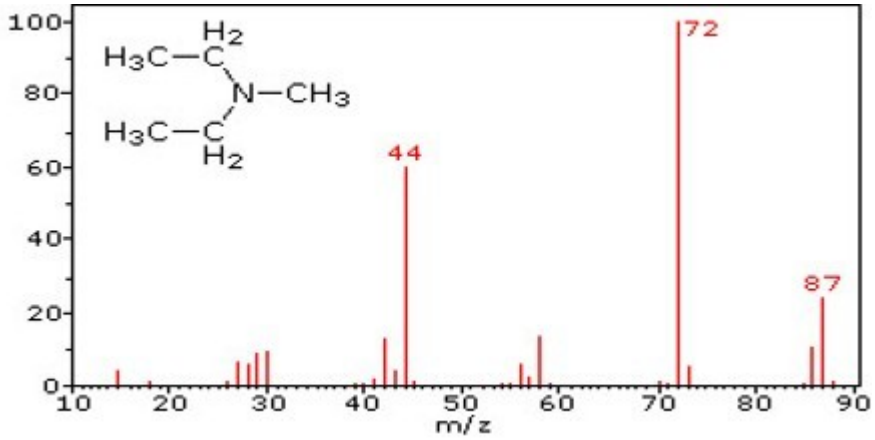
Class so far...

- Deterministic searching (counting, clumps, KMP)
- Randomized searching (Gibbs sampling)

- This week: Branch and bound search

Mass Spectrometry





Random breakage

Glu Leu Val Ile Ser Ile Ser Ala Leu Ile Val Glu

ELVISISALIVE weight (ELVISISALIVE)

E LVISISALIVE + weight(E) + weight (LVISIS...)

EL VISISALIVE + weight (EL) + ...

ELV ISISALIVE

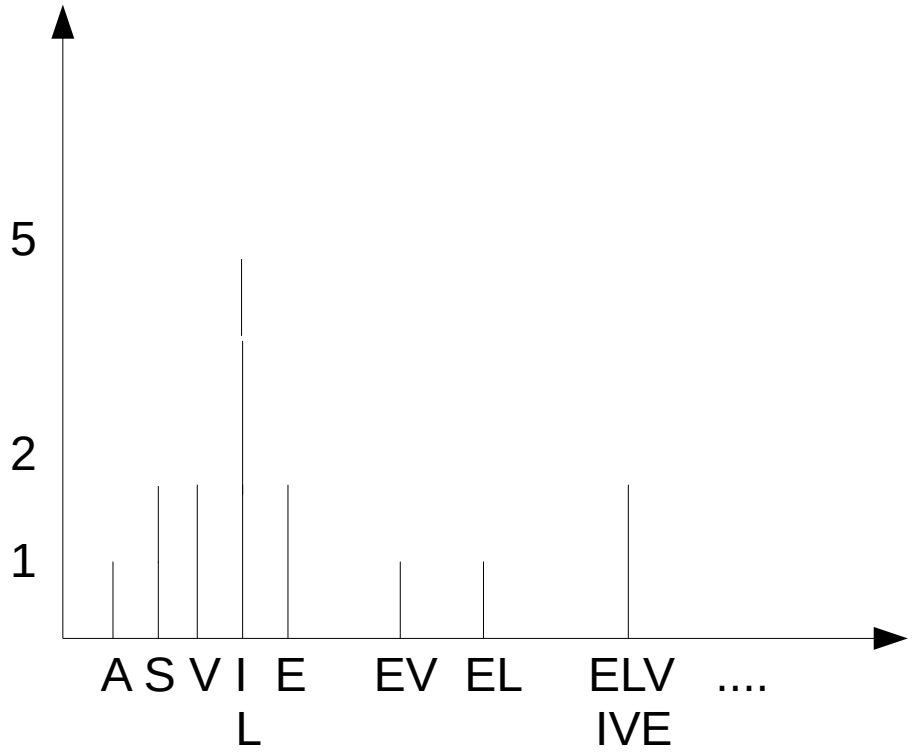
ELVIS ISALIVE

ELVISI SALIVE

ELVISIS ALIVE

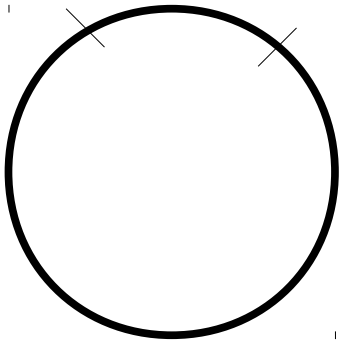
...

Peptide spectrum



Cyclic spectrum?

Exactly 2 cuts



ELVISISALIVE

E LVISISALIVE
L VISISALIVEE
...
ELVI SISALIVE
LVIS ISALIVEE
...
SALI VEELVISI
...

Whole peptide + all possible breaks into 2 pieces

What is the runtime for creating the cyclic spectrum of a peptide of size length k?

Our goal

- Given a real spectrum
- Find peptide that generated it

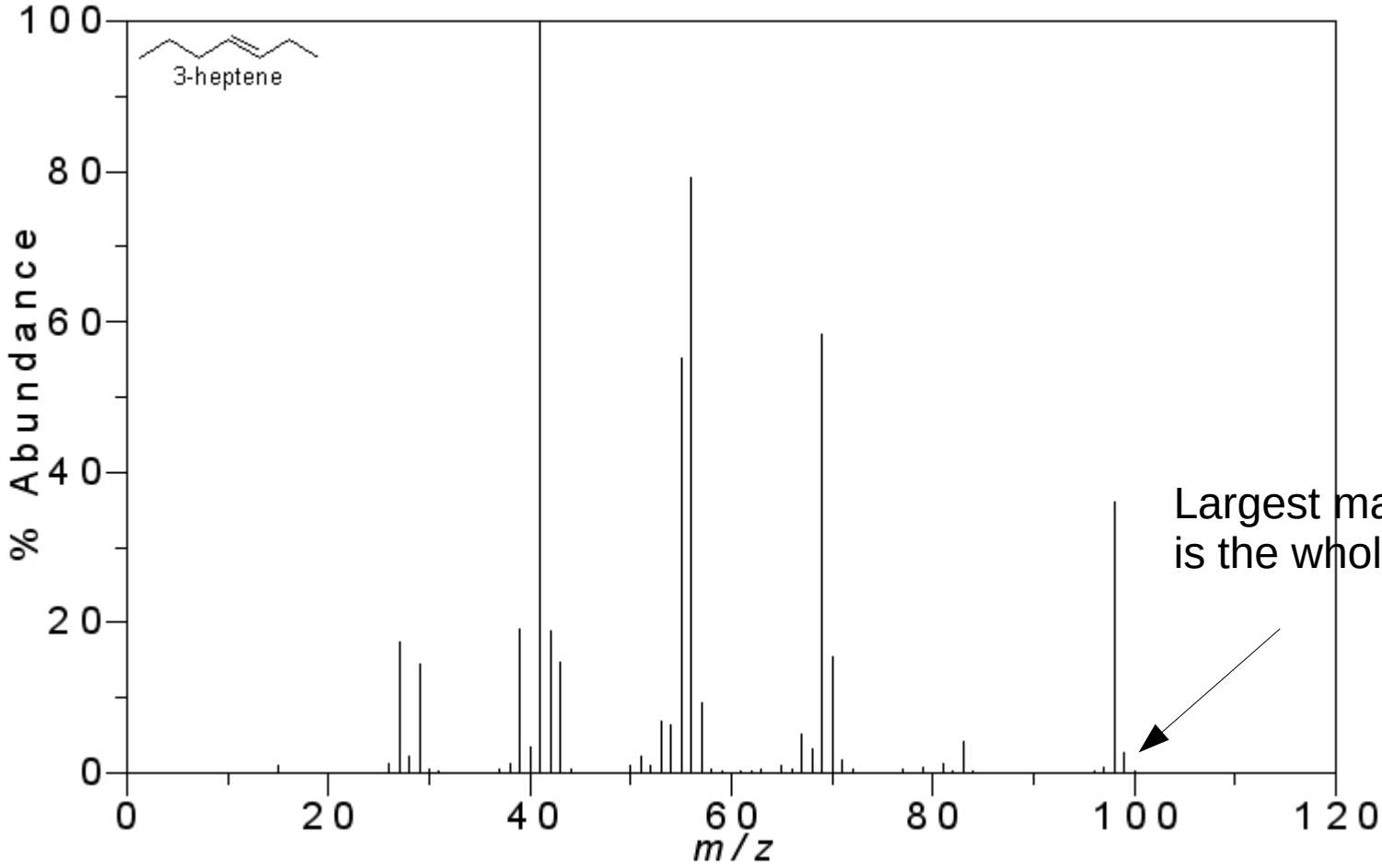
- The other way around is easy:
  - break peptide in every each way
  - calculate weights
  - compute predicted spectrum

General approach

- Guess a peptide
- See how well it matches spectrum
- Come up with new guess
- Repeat

- Sound familiar?

Some quick insights



How many peptide words have a given mass?

How many combinations of coins and bills make a same $ amount?, e.g., $0.5

brute force – not so simple

1. 1 + 1 + 1 + 1 + ... + 1 = $0.50
2 . 1 + 1 + 1 + 1 + ... + 5 = $0.50
3. 1 + 1 + 1 + 1 + ... + 5 + 5 = $0.50
...
11. 5 + 5 + 5 + 5 + ... + 5 = $0.50
12. 5 + 5 + 5 + 5 + ... + 10 = $0.50
...
16. 10 + 10 + 10 + 10 + 10 = $0.50
17. 1 + 1 + 1 + 1 + 1 + 5 + ... + 5 + 10 = $0.50
...

MANY!

weight(ELVISLIVES) = weight(ELVESSLIVI)

i.e., need to take into account the whole spectrum!!

Key insight

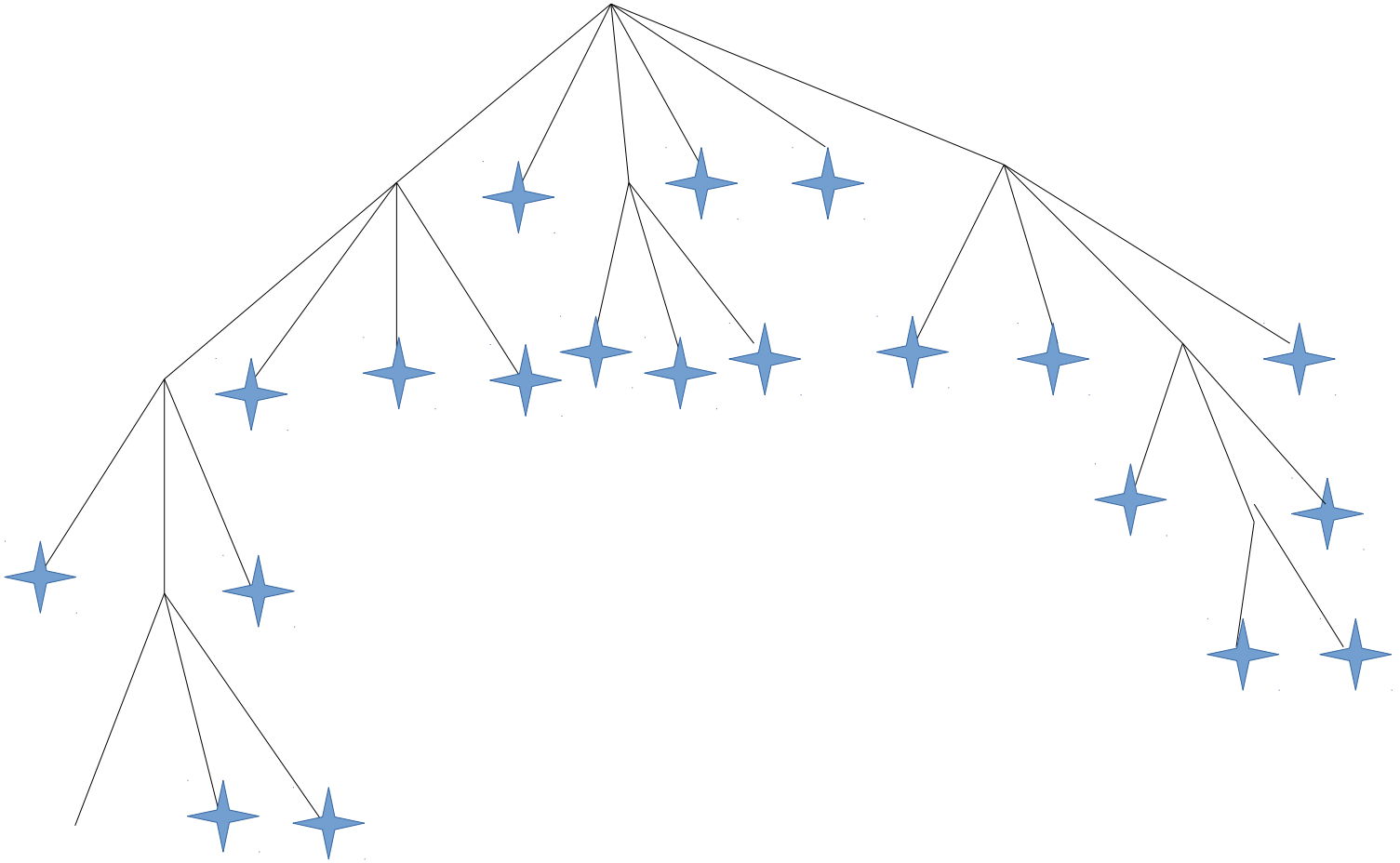
- Bad guesses have bad cyclic spectra

- LIVES, IVES, IVE
  - are in spectrum(ELVISLIVES)
  - but not in spectrum (ELVESSLIVI)

Algorithm 1

- Assume experimental spectrum is perfect
- Generate all peptides of length 1
- Discard the ones not found in spectrum
- Extend the remaining ones by one amino acid
- Discard the ones incompatible with spectrum
- Repeat...

Branch and bound



Without the "bound" step, how big is the search space for a peptide of length k?

Spectrum "matching" algorithm?

Dealing with errors

- Even one error in experimental spectrum can "disqualify" correct answer
- Remind you of anything you've seen?

- Instead of "match/no match" look for score of match: # of masses in theoretical spectrum found in experimental spectrum

- Why not also account for # of masses in experimental spectrum not found in theoretical spectrum?

New algorithm

- Assume experimental spectrum is perfect
- Generate all peptides of length 1
- Keep the best matching one
- Extend it by one amino acid
- Keep the best matching one
- Repeat...

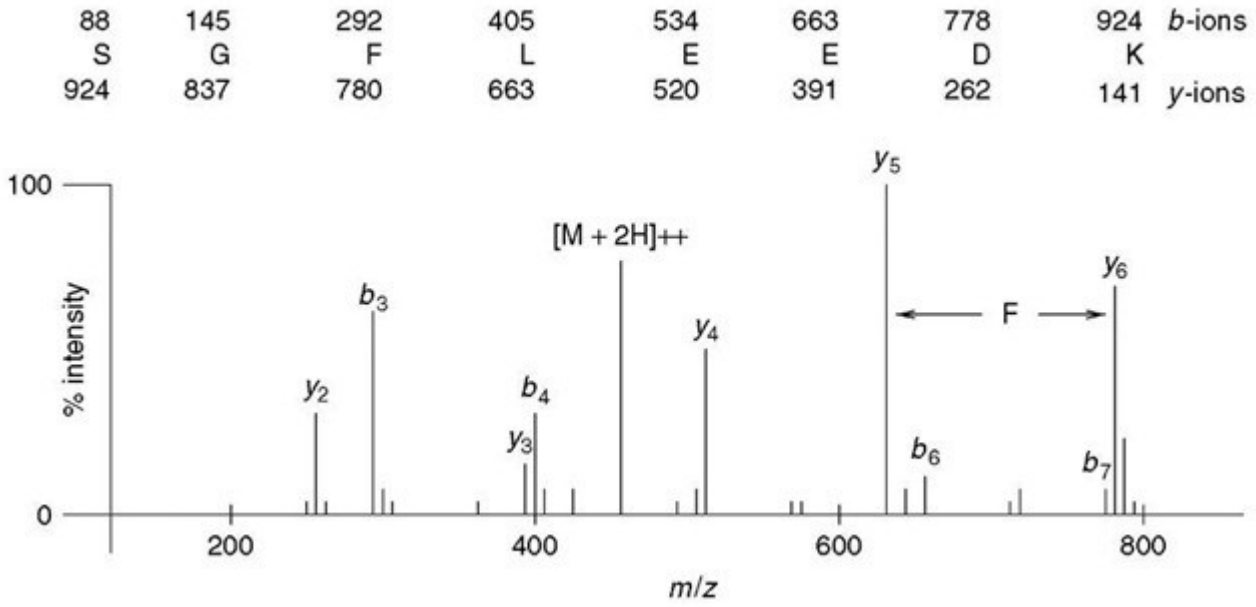
Any issues?

New algorithm

- Assume experimental spectrum is perfect
- Generate all peptides of length 1
- Keep the best matching ones (top of the leaderboard)
- Extend it by one amino acid
- Keep the best matching ones
- Repeat...

What if you don't know weights?

- Easy – infer from experimental spectrum



ELVISLIVES

E = ELVISLIVES – LVISLIVES
E = SELVISLIVE – SELVISLIV
E = ELV – LV
E = LIVE – LIV
...

The most frequent small differences are the amino acid masses

Full algorithm

- Infer amino-acid masses from spectrum (if you cannot trust your database)
- Run leaderboard algorithm

- Will this stop at some point?